Abstract: Data sharing is in this report understood as the release of data for use by others for the purpose of learning analytics. The issues for data sharing identified through case studies and other means are analysed from a technical, semantic, organisational, and legal perspective, following the structure of the European Interoperability Framework. The issues are also assessed as to whether they could easily be addressed or would appear as a blocker of progress. Even if the research community is rather optimistic of the possibility to “scale up” learning analytics to create big impact using big data, this report shows the need for more research and community exchange to frame, sort and prioritise the issues of data sharing.
Executive summary

Data sharing is in this report understood as the release of data for use by others for the purpose of learning analytics (LA). We are not exploring sharing of data between software under control of a single legal entity, but data sharing between persons or organisations.

Data sharing is a precondition for realising the potential of learning analytics. Even so, it is a finding in this report that the ramifications of data sharing have not been worked out, and that there is work to be done to characterise the problem-space, as well as the solution-space.

Through exchange with the learning analytics research community and collection of case reports from a number of stakeholders in schools, universities and industry, we have gathered requirements for data sharing in order to explore the feasibility of actions within a number of perspectives.

The variety of data that is relevant to learning analytics is very great, and in exploring aims for data sharing it is useful to discern between data sharing benefitting analytical and research processes, and data sharing supporting and enhancing the organisational and pedagogical processes contributing to the learning outcome. Eight aims were identified and assessed towards the current interest of the LA stakeholder communities.

- More useful analysis through the combination of data from different sources
- Sufficient scale of data to determine relevance and quality of educational resources
- A critical mass of data for learning science research
- Reproducibility and transparency in learning analytics research
- Cross-institutional strategy comparison
- Research on the effect of education policy
- Social learning in informal settings
- Learner data as a teaching and learning resource

The issues for data sharing identified through case studies and other means are analysed from a technical, semantic, organisational, and legal perspective, following the structure of the European Interoperability Framework. The issues are also assessed as to whether they could easily be addressed or would appear as a blocker of progress. The following issues are identified within each category:

Technical and semantic issues
- Incompatible technical implementations
- Inappropriate architectural assumptions
- Inadequate domain-level semantic models
- Lack of adoption of existing specifications

Organisational issues
- Privacy
- Inadequate decision-maker knowledge

Legal issues
- Lack of awareness of what is possible within the law
- National differences
- Current legislation is out of date
The report also discusses how the identified issues could be addressed and how feasible their solutions are. Again the European Interoperability Framework is used as a scaffold. These are the activities that could be undertaken, and that are classified as being either challenging, requiring concerted efforts, or are seen as uncomplicated:

**Technical and semantic issues**
- Shared open architectures and common frameworks
- Code-bashes (plugfests) - addressing mid-level practical interoperability
- Practice-oriented pre-standardisation at the domain-level

**Organisational issues**
- Anonymisation and statistical disclosure control
- Analytics models as shared data
- Remote access analytics
- Trusted data analysts
- User-managed access
- Common codes of practice and standardised data agreements
- Develop understanding and consensus around risk-based approaches to privacy protection

**Legal Issues**
- Raise awareness of what is possible within the law

Even if the research community is rather optimistic of the possibility to “scale up” learning analytics to create big impact using big data, this report shows the need for more research and community exchange to frame, sort and prioritise the issues of data sharing. Moving towards a LACE Data Sharing Roadmap this report organises 11 issues on a time scale of action, and groups them according to type of issue. There are a number of issues that can be addressed right now. Other issues need more time for stakeholders to take action, while many contextual issues need to be further explored to understand how they could be addressed.

This report is based on desk research by the LACE Work Package on Interoperability and Data Sharing and case studies with contributions from stakeholders and other EU projects. The case studies are included in Part B of this report.
## Contents

Guide to this document ................................................................................................................. 1

1 Introduction to data sharing for learning analytics ................................................................. 4
  1.1 Big data or small data ........................................................................................................... 4
  1.2 Data and people – the wider social and technological landscape ...................................... 5
  1.3 Definition of scope .............................................................................................................. 6

2 Exploring aims for data sharing .............................................................................................. 8
  2.1 Eight aims for learning analytics data sharing ................................................................. 8
  2.2 Rating of aims ................................................................................................................... 11

3 Issues for data sharing ........................................................................................................... 13
  3.1 Technical and semantic issues .......................................................................................... 14
  Summary technical and semantic issues ................................................................................. 16
  3.2 Organisational issues ....................................................................................................... 17
  Summary organisational issues .............................................................................................. 21
  3.3 Legal issues ....................................................................................................................... 21
  Summary legal issues ........................................................................................................... 22

4 Addressing issues .................................................................................................................. 24
  4.1 Addressing technical and semantic issues ....................................................................... 24
  Feasibility of actions ............................................................................................................... 26
  4.2 Addressing organisational issues ....................................................................................... 26
  Feasibility of actions ............................................................................................................... 31
  4.3 Addressing legal issues ...................................................................................................... 31
  Feasibility of actions ............................................................................................................... 32

5 Roadmap .................................................................................................................................. 33
  5.1 Business and political context .......................................................................................... 33
  5.2 Concerning actors .............................................................................................................. 35
  5.3 Suggested steps to take ..................................................................................................... 37

6 Conclusion ............................................................................................................................... 39

PART B 1 LACE Workshop on Privacy and Ethics ................................................................. 41
PART B 2 European Project Case Studies .............................................................................. 44
  2.1 Designing a national learning analytics system – the case of Estonia ......................... 44
  2.2 PELARS – a European project investigating practice-based learning ......................... 48
  2.3 LEA’S BOX – a European project creating a Learning Analytics Toolbox ................... 50
  2.4 WATCHME project – case ............................................................................................... 52
2.5 Building a cross-sector service platform – the case of Connect, Norway ............................. 53
2.6 Requiring an open architecture for learning analytics – the case of JISC plans for the UK ...... 55
2.7 A Norwegian learning analytics vendor’s perspective on data sharing and interoperability .... 59
2.8 A Norwegian publisher with adaptive learning product cooperating with a US company ........ 61
2.9 School owners’ concerns – what is allowed according to the Privacy Protection Act? The case of giving advice to Norwegian schools .................................................................................................................. 63

Part B 3 Existing examples of data sharing for learning analytics ............................................. 65
3.1 Pittsburgh Science of Learning (PSLC) DataShop ................................................................. 65
3.2 Predictive Analytics Reporting Framework (PAR) ................................................................. 67
3.3 Open Academic Analytics Initiative (OAAI) ......................................................................... 68
3.4 Multimodal contextualised Learner Corpus Exchange (MULCE) .......................................... 69
3.5 Stanford Data Portal for Research (VPOL) .......................................................................... 71
3.6 xMOOC platforms ............................................................................................................... 72
3.7 Open data initiatives ........................................................................................................... 73
3.8 A counter-example: InBloom .............................................................................................. 74

Glossary of abbreviations and specialist terms ........................................................................ 76
References ................................................................................................................................ 77
About ....................................................................................................................................... 79
Guide to this document

This report is the LACE Project’s contribution to developing an appreciation of data sharing - the release of data for use by others - as an umbrella idea that covers a range of possibilities, within the field of learning analytics. We consider both data sharing for learning analytics in practical teaching and learning situations and for learning science research, but rather than focus on one motivation for data sharing, we try to capture variety within the field.

Part A of this report is concerned with the scope of data sharing for learning analytics, the motivations for doing so, and with the issues involved and ways in which these issues may be overcome or navigated-around. Part B presents a number of case studies and examples which complement the desk research used to create Part A. Forward references are made from Part A to Part B to make links to evidence rather than to direct the reader to an elaboration of the point; Part B is a supplement.

Part A is structured as follows:

- Section 1 provides some of the contextual background to data sharing for learning analytics and outlines the scope we have adopted for this report.
- Section 2 presents a number of aims for data sharing, covering both research and operational use of learning analytics. The section concludes with a rating of aims according to our assessment of learning analytics community interest.
- Section 3 explores the issues arising from data sharing, obstacles to realising the aims. Issues of diverse character are considered: technical, organisational, legal, etc. The section concludes with a mapping of the significance of each issue to the aims identified in subsection 2 in order to arrive at a rating of the extent to which the issue is an obstacle to the progress of learning analytics in general, i.e. how important the issue is.
- Section 4 discusses a variety of ways in which the issues identified in subsection 3 might be addressed. Some actions that would be required to make progress are advanced and classified according to broadly-defined levels of feasibility.
- Section 5 concludes Part I with a look at the intersection of importance of issues and the feasibility of ideas to address the issues. This leads to an outline roadmap, which is supplemented by a discussion of the potential roles of change-agents in the field of learning analytics.

Whereas Part A of this report is concerned with the scope of data sharing for learning analytics, the motivations for doing so, and with the issues involved and ways in which these issues may be overcome or navigated-around, Part B provides background material which we have combined with desk research to create Part A. This evidence is largely presented in this document and is organised into three subsections:

- Section 1 presents the views of participants in a series of LACE workshops on privacy and ethics.
- Section 2 comprises a set of case studies of a selection of European learning analytics initiatives in the public and private sectors. These have been obtained from, or in collaboration with, third parties. We have added a boxed summary of our observations which draw out from each unit the significant points relevant to this report.
• Section 3 presents a summary of operational learning analytics data sharing initiatives. This is an extended excerpt from a component - “Learning Analytics Data Sharing - Current Examples 2014” - of D7.1 (Cooper & Hoel, 2014).

In addition to this evidence, Part A has drawn on the Learning Analytics Specifications and Standards Quick Reference Guide, which was included as a snapshot in D7.1 and which is not duplicated here.
Part A
1 Introduction to data sharing for learning analytics

1.1 Big data or small data

Learning analytics processes necessarily involve data, and once you start talking about data, the Big Data meme pops up and brings images of gigantic amounts of digital data controlled by companies, authorities and large organisations, subject to extensive analysis based on the use of advanced algorithms.

To create a feeling of urgency and change when talking about big data and analytics one is often presented with the fact that over 90 percent of the world’s data was produced just in the past two years (Gudivada et al., 2015). Data related to learning is just a tiny fraction of the stream of data emanating from nearly all aspects of our daily life. The claims for the promise of Big Data are far reaching, driving “more accurate descriptive and predictive models that inform decision making on every level, whether identifying the next big security threat or making the best diagnosis and treatment choice for a given patient”, observes (Schadt, 2012). The influence of Big Data is considerable for a wide range of life spheres and these experiences will collectively shape the way we are willing to share our personal data in general. As a politician said during a LACE policy brief, if Big Data has revolutionised advertising, it can certainly help us improve education! But is the archetypal corporation with Petabytes of data a suitable prototype for learning analytics?

The situation for educational institutions is more characterised by Small Data than Big Data and there are certainly differences between selling soap and grafting knowledge, including many subtle aspects and contested notions about purpose, methods, and values. It is often said in workshops on learning analytics: currently it is not so much about Big Data, – you can easily fit your data in a spreadsheet on your laptop computer! While a spreadsheet may be inadequate for captured activity data, it remains true that for most of the potential applications of learning analytics in education and training practice, the useful data will be of a scale well below that of Big Data.

Many applications of learning analytics are found to require data sharing to realise their potential. For example, large scale data is often a prerequisite for educational data mining techniques or multivariate statistics. Although the data from an institutional learning platform or a MOOC may be considered large and varied, the scale and coverage of such datasets may be insufficient to allow the potential of learning analytics to be fully realised because of the great variety of learner and contextual attributes. For example, Verbert et al. (2011) demonstrated that the activity of learners on a single course is likely to be so diverse that a learning resource recommender system would be practically useless if only based on data at this scale. Thus, this challenge applies to both learning science research and to potential products and services built around data generated during learning activities. It is also usually the case that the data required to undertake learning analytics resides in different software systems. Increasing use of Cloud Computing models of service and IT provision, where expertise or technology is provided by a separate organisation to the education provider, has increased the extent to which data is not only distributed between different IT systems, but is also distributed among legal entities. The combined effect of the need for data at scale and for combination of different data sources motivates the idea that data sharing between organisations - potentially including public and private sector bodies - is an important enabler for effective learning analytics.
A distinctive difference between the archetypal Big Data corporation and educational establishments is, therefore, the requirement for data sharing. During the preparation of this report, the authors have been struck by the extent to which the characterisation and ramifications of data sharing have not been worked on. Although we have found a number of examples of data sharing, we have also discovered that conversations about data sharing with various stakeholders have often raised questions and sometimes exposed un-asked questions that should be addressed. There is, it seems, some work to do to properly characterise the problem-space, as well as the solution-space.

Whatever we would like to do in relation to data sharing for learning analytics will be moulded by the impact of trends outside of education. While the education community is likely to place barriers to prevent data from the outside world from being used in learning analytics (The Open University 2014a, 2014b), we also know that that private space is already being invaded by sensors and data hungry applications, with personal information frequently shared to interests pursuing less noble aims than promoting the individual’s knowledge and skills. In only a few years, we have moved to the point where GPS, accelerometers, and heart rate sensors are common-place, in which location data is leaking out through mobile apps, and technologies such as iBeacon are being launched which tighten the connection between location and retailing in non-transparent ways (Scoble & Israel, 2014).

The boundaries between the worlds of formal and informal learning have always been somewhat permeable and ill-defined; and in world in which numerous digital devices can capture and render any activity or utterance into data it would hardly be practical to isolate education for the “outside”. It will be difficult to raise higher standards for personal protection within education than in other domains. Before focussing on the learning analytics context, therefore, it may be useful to step back and to consider the wider social and technology landscape and to give some definition to our scope of “data sharing”.

### 1.2 Data and people – the wider social and technological landscape

In the book “Age of Context: Mobile, Sensors, Data and the Future of Privacy” (2014) the technology journalists Robert Scoble and Shel Israel examine five forces creating a storm of change: mobile devices, social media, big data, sensors and location-based services. Learning analytics is an important, but only a small part of this storm. When controversies build around learning analytics specific issues like access to social media data, learner control of information, parents’ rights to follow each step of the educational process, etc. it is important to realise that learning analytics shares challenges with other sectors and that more and more of our society nowadays is driven by data. The solutions found by the learning analytics community will be impacted by the development in the society as a whole. Just reflect on the trends that underlie future development in the standards community, captured by these headlines from a recent technology trend:

- Computing Everywhere;
- Advanced, Pervasive and Invisible Analytics;
- Social Business with a special focus on Social Network Analysis, Social Content, and Social Commerce;

---

2. see the Joint Technical Committe 1 by ISO: [http://www.iso.org/iso/jtc1_home.html](http://www.iso.org/iso/jtc1_home.html)
• Risk-Based Security and Self-Protection;
• Wearables electronics and services; and
• Privacy Management with a special focus on Privacy Management Tools.

These trends will also impact learning, education and training.

When analysing data sharing for learning analytics, the subject of this report, we need to identify the specifics of learning analytics while making sure that we understand the dynamics of a field that is entangled in a more complex system where solutions to challenges and barriers may be sought in knowledge domains not usually visited by the educational profession, like health informatics or cyber security (section A 4.2 Addressing organisational issues).

1.3 Definition of scope

1.3.1 Interpretation of “data sharing”
Data sharing can be defined as the release of data for use by others. For the purpose of this report, we constrain the term to refer to data sharing between legal entities (an organisation or person), and not just sharing between software under control of a single legal entity. This outlook on data sharing requires careful consideration of factors which do not apply to the single-operator system-integration scenario. It is these factors, rather than the details of the system-integration problem, which are the focus of this report.

The interpretation of data sharing is at present somewhat confused by common conceptions of ownership and related factors, under-developed thinking about the topic, and sometimes a failure to consider it. The increasing use of software hosted in “the cloud” - i.e. Software as a Service (SaaS) - has amplified this situation, but we have also inherited a confusion from the days when most software used in education was running on-site. This confusion relates to ownership and control, and the extent to which the educational establishment has absolute authority or acts as a custodian on behalf of the learner. While it is clear that there is some data which the educational establishment is required to keep, and some of which the learner has no right to change, there has generally been little attention to the details of ownership, control, and custodianship. The rise of SaaS has introduced the third party service providers into the picture, and many of these provide at-best limited access to the data created through the use of their software. The rise of interest in learning analytics is starting to make the situation better known, but procurement specifications and service level agreements continue to lack clarity on the distinctions that can be made in ownership and control, and tend to presume absolute power resides in the service provider. In many cases, this situation makes for vagueness when considering who would release data and to whom. A consequence of this situation, which constitutes a background issue for making progress with learning analytics data sharing, is that this report will discuss “data sharing” with some generalised assumptions about who has control, and who will use the data.

When sharing research data a systematic approach for deposit, sharing, reuse, curation and preservation of data is required (van den Eynen & Bishop, 2014). All these processes are definitely present also in a learning analytics data sharing scenario; however, the requirements may differ. For instance, preservation of data may be a more pressing need for sharing of research data than for learning analytics data. Learning is after all a more transitional activity, where even the right not to have one’s learning trajectory stored may be in the interest of the learner.
1.3.2 Types of data
The variety of data that is relevant to learning analytics is potentially very great indeed. For the purpose of this report, our interest is primarily in data about people or their activity in a learning-related situation. We are not concerned, for example, with national or international classification schemes for subject matter of courses or learning resources. These are not our focus of attention for two reasons: sharing such reference data lacks many of the complications of person-related data; the sharing of this kind of impersonal data of utility for learning analytics is often undertaken as an Open Data initiative and recent projects, such as Linked-Up\(^3\) have produced thorough accounts of that topic. Consequently, we pay little attention to Open Data in Part A, although a short section in Part B is devoted to the topic (see B 3.7 Open data initiatives). In short: a focus on person-related data gives us more clarity of purpose because this relationship influences so many critical decision-points.

---

\(^3\) [http://linkedup-project.eu/](http://linkedup-project.eu/)
2 Exploring aims for data sharing

This section is written not as a definitive guide to, nor survey of, the aims and motivations for data sharing for learning analytics but as an exploration of the range of possibilities. Some of these are drawn from the examples and case studies which appear in Part B, while others have been stimulated by workshop sessions and informal conversations at, for example, LAK 2015 (including the Open Dashboard Hackathon), EC-TEL 2014, the Open Learning Analytics Summits, the EP4LA workshops, the Jisc co-design workshop on effective learner analytics, etc. These have involved researchers, HE managers, and IT specialists.

In respect of aims, we wish to be clear that we do not understand “learning analytics data sharing” to represent a single idea. Similarly, even within our chosen scope of person-related data, questions about privacy, feasibility, sustainability, etc. vary greatly depending on whether the data in question is about records of education achievement and history, public social media activity, LMS/VLE activity, etc. Consequently, it is clear that there cannot be a single idealised technology platform, workflow, policies, etc.

In exploring aims for data sharing we have considered both: the affordance of data sharing to the analytical and research processes; and the benefits to learning outcomes from enhancement of organisational and pedagogical processes, where data sharing is a precursor to informed process change. In crude terms, some aims are related to research and others to practice, but this does not imply we see a separation of research and practice as being necessary or desirable. Indeed, it is arguable that a better connection between research and practice is desirable, and this is attested to by the activities of SoLAR\(^4\) in seeking to engage more practitioners. It is conceivable, for example, that a data sharing platform intended to primarily serve the enhancement of outcomes would be a valuable research resource, and that concurrent use for academic research and organisational and pedagogical processes would catalyse cross-over and synergetic projects.

2.1 Eight aims for learning analytics data sharing

Data sharing aims to support learning analytics on different levels, from contributing to new knowledge and research through bringing more data for research and development; improving organisational analysis through aggregations of educational data to improve retention strategies and learning design; to active learning support through synchronous adaptation of the learning context and content. The following eight aims capture some, but certainly not all, of these complex requirements for data sharing:

**Aim 1: More useful analysis through the combination of data from different sources**

It is the norm, rather than the exception, for learners to engage in activities in many different pieces of software. In this case, the value that can be gained by treating each data-set independently is quite low, and many important questions can only be addressed with a more complete picture (we recognise that many questions remain unanswerable from captured data alone). This aim is relevant to all outcome categories. Were all of these software systems to be operated by a single educational establishment, the challenge would be largely one of system integration. This scenario is, however, rare, even for large universities with access to skilled developers and infrastructure engineers. Aim 1 encompasses scenarios where software is operated by third parties on behalf of an educational establishment, even for large universities with access to skilled developers and infrastructure engineers.

---

\(^4\) [http://solaresearch.org/](http://solaresearch.org/)
establishment and scenarios in which learners engage in learning-related tasks using their personal social web services. It will be important to tease-apart these scenarios when considering the issues raised by data sharing. Whereas Aim 1 is concerned with variety of data, Aims 2 and 3 are concerned with its scale.

**Aim 2: Sufficient scale of data to determine relevance and quality of educational resources**

This aim would be shared by the creators of resource recommender systems, or for more open ended analysis of learner behaviour in an institutional context (e.g. to for learning design, learning support, or resource planning). Publishers may also wish to use this data for improving their products. In practice, there is likely to be a strong link to Aim 1: data on usage within the nominal resource - e.g. down to page or lower level in a text - or on actions such as annotation, when joined with other learner data offers substantially greater benefits than anonymous usage would.

**Aim 3: A critical mass of data for learning science research**

This aim is a counterpart to Aim 2, tuned to research process rather than education/training outcomes. Within Aim 3, two aspects are combined: the need for large scale data, and fact that the effort/cost of acquiring high quality data is a limit to the pace of research. This aim fundamentally relates to increasing the quality, or decreasing the cost, of pure research outcomes. The examples in sections B 2.2 PELARS – a European project investigating practice-based learning, B 3.1 Pittsburgh Science of Learning (PSLC) DataShop, B 3.4 Multimodal contextualised Learner Corpus Exchange (MULCE), and B 3.5 Stanford Data Portal for Research (VPOL) illustrate this aim.

Data at scale may be used to overcome challenges that have been levelled at conventional forms of educational/psychological research, in which reduction of the problems to a form that is amenable to small-data research introduces mis-specification into the research question (Winne, 2006). In brief, since educational outcomes, and related aspects of interest to various stakeholders, such as employability, depend on a large number of parameters, a large volume of data is required to address many research questions in education. This is likely to mean both data from a large number of individuals as well as a substantial quantity of data from each. There is, therefore, a limit on what can be achieved using data from a single cohort or educational establishment. The same benefits also apply to the development of educational technologies.

The desirability of sharing for data re-use, to increase the ratio of data use over data gathering, is taken as not requiring explanation.

**Aim 4: Reproducibility and transparency in learning analytics research**

Both the process and outcomes perspective stand to benefit from wider access to the data used in research work, whether this is overtly academic research or overtly institutional research. This could include two related activities: reproducible research, and the re-interpretation of findings.

[Aim 4a] Reproducible research refers to research in which data and computer code (etc.) are made available in addition to published papers, such that other researchers can reproduce, rather than attempt to recreate, the procedure that led to the paper. This is a research process aim. Interest in

---

5 The term “institutional research” is borrowed from North American usage. It should be understood, in this document, to indicate research activities that are undertaken with the intention of improving organisational or pedagogical processes rather than for the pursuit of knowledge per se. It is being used to identify a subset of learning analytics and should not be interpreted as meaning institutional research other than of a learning analytics character; we are not introducing ostensibly management interests.
reproducible research has grown from a desire to: deal with issues of efficiency in a research community, where effort to verify or adapt research, including the conduct of meta-analysis and comparative studies, is currently often substantial; to increase the extent to which research is a public good, for example through Open Access publication. FOAS, the Foundation for Open Access Statistics, for example, has a mission to “promote free software, open access publishing, and reproducible research in statistics”. The Journal of Learning Analytics (JLA) is already an Open Access journal and a number of participants in the Open Learning Analytics Summit (Indianapolis, March 2014) advocated support for reproducible research as being something JLA should move towards.

[Aim 4b] The re-interpretation of findings kind of activity might fall short of full reproducibility but would allow claims for efficacy to be subject to inspection and counter-interpretation. This may be more relevant for institutional research or to determine the effectiveness of learning analytics products, where there is a wider community interested in verifying and generalising what works in one institution to other institutions with similar contextual factors and student profile. The need to do this is becoming more widely felt as the practical difficulties in demonstrating “what works”, and the contextual prerequisites for it, are being repeatedly demonstrated. Education and educational processes are complex and frequently wicked\(^\text{6}\). The challenges raised against Purdue University over its published claims of efficacy of Course Signals, and the controversy that followed\(^\text{7}\), illustrates a scenario which this aim could avoid.

**Aim 5: Cross-institutional strategy comparison**

There is considerable interest, both in educational establishments and in public-sector agencies in understanding how well different strategies for teaching and learning, and the support of teaching and learning, perform and the extent to which these represent sharable “good practice” in similar educational and cultural settings. These would include “intervention strategies” to reduce drop-out or under-achievement, for example, but assessment strategies and investigation of strategies such as “flipped classroom” fall under the same category. The core of this aim is the comparison of the effects of strategy, but this is unlikely to be a meaningful comparison if summative “benchmark” statistics are shared, although sharing a wider range of aggregated outcomes than at present would achieve some benefits. More meaning, and understanding of what it is about a strategy that makes it work and in which contexts, is likely to require sharing of more detailed data. There may be both organisation-level and person-level concerns about who can access this data. The PAR Framework (section B 3.2 Predictive Analytics Reporting Framework (PAR)) illustrates this aim.

**Aim 6: Research on the effect of education policy**

Education ministries, and local authorities, have used data for this purpose for decades, but the data they have used has often been quite crude and there have usually been long time lags. This state of affairs reflects a heritage of labour-intensive reporting processes, which Civil Servants and Ministers are now challenging and beginning to express an interest in data that is more detailed and closer to

\(^{6}\) “wicked” is used in the sense of “wicked problem”, which Wikipedia summarises as: “a wicked problem is a problem that is difficult or impossible to solve because of incomplete, contradictory, and changing requirements that are often difficult to recognize.” ([http://en.wikipedia.org/wiki/Wicked_problem](http://en.wikipedia.org/wiki/Wicked_problem))

“live data” (Plant, 2013). This aim is a close relation to Aim 5, but separated because of the different locus of control. In reality, both the policy environment in which educational establishments operate, and their locally-selected (or developed) strategies are entangled. The case study in B 2.1 Designing a national learning analytics system – the case of Estonia illustrates this aim.

Aim 7: Social learning in informal settings
This aim is unique among all that we outline in that it is concerned with autonomous sharing of learning-related data by learners with self-selected peers. We have included it not as a claim that this is a well-established aim for data sharing, but to draw attention to a possibility that has been largely neglected in the learning analytics discourse so far, although some initiatives are interested in the idea of personal data stores (e.g. see the Estonian case study, section B 2.1 Designing a national learning analytics system – the case of Estonia). The essential idea is that sharing and comparing data with a peer has the potential to significantly expand on the benefits to metacognition and self-efficacy, etc., relative to isolated introspection of a “quantified learner”.

Aim 8: Learner data as a teaching and learning resource
Analysis of learner data is likely to become a growing component of undergraduate and masters-level teaching in education, psychology, and machine learning courses. This assertion is evidenced by the announcement from the Society for Learning Analytics Research that it intends to collaborate with the International Educational Data Mining Society8 on an initiative to create an open masters program to be licenced for use and re-use by academic institutions, the Learning Analytics Masters Program. Authentic data with known data quality and suitable for teaching specific principles and processes would be a valuable asset, and the potential levels of use could be many orders of magnitude greater than for Aim 4.

2.2 Rating of aims
In order to guide the identification of issues, which are discussed in section A 3 Issues for data sharing, to be estimated, we will organise the aims according to our assessment of learning analytics community interest using a four level ranking (see Table 1).

1. Speculative
2. Plausible mid-term or minority interest
3. Near-term and broad interest
4. Here now

The assessment of interest is our subjective combination of intelligence from: the case studies and examples in Part B, academic and non-academic conferences and workshops, grey literature, standardisation project teams, etc. The rating numbers indicate ordering; we do not imply proportional weighting.

<table>
<thead>
<tr>
<th>Aim</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 – Speculative</strong>&lt;br&gt;Aims which may be identified only by a few people in the wider learning analytics community and which may well remain that way</td>
<td>Aim 7: Social learning in informal settings&lt;br&gt;While social and informal learning is widely appreciated in general, the idea of peer-to-peer data sharing for learning analytics is</td>
</tr>
</tbody>
</table>

---
## Marginal interest.

### Level 2 – Plausible mid-term or minority interest

*Aims which would be identified only by some groups within the learning analytics community but which may plausibly become of wider appeal in the mid-term*

| Aim 4a: Reproducible Research | While reproducible research has a number of purposeful proponents, and could have a dramatic effect on scholarship, it is an idea which would require explanation to most learning analytics stakeholders. |
| Aim 8: Learner data as a teaching and learning resource | The inevitable rise in courses with an overt learning analytics component, or of data handling in general, will surely lead to keen interest in suitable data. |

### Level 3 – Near-term and broad interest

*Aims which are likely to be identified by large sections of the learning analytics community, and for which the benefits are appreciated*

| Aim 2: Sufficient scale of data to determine relevance and quality of educational resources | While educational publishers, and some content brokers, are doing this now, the extent to which data sharing is occurring (and having benefits) is limited. |
| Aim 4: Transparency in the success of learning analytics products and initiatives | As more institutions move to adopt learning analytics products and processes, we can expect greater scrutiny of what works, and the contextual factors which influence efficacy. |
| Aim 5: Cross-institutional strategy comparison | The PAR Framework (section B 3.2 Predictive Analytics Reporting Framework (PAR)) includes this now, but the idea of institutions using analytics to compare each other’s student support strategies is yet to spread. |
| Aim 6: Research on the effect of education policy | The use of more fine-grained educational data than has conventionally gathered as government statistics seems to be a widely appreciated idea, but one which is taking time to translate to practice. |

### Level 4 – Here now

*Aims that are being pursued by a number of organisations at present*

| Aim 1: More useful analysis through the combination of data from different sources | This aim is evident in the many projects seeking to capture data from multiple sources into some form of data warehouse using technologies such as IMS LTI and ADL xAPI. |
| Aim 3: A critical mass of data for learning science research | As indicated in the outline of Aim 3, and in section Part B 3 Existing examples of data sharing for learning analytics, there are several examples of initiatives focussed on learning science research. |

**Table 1: Rating of the aims**
3 Issues for data sharing

Issues for data sharing may be looked at from a range of perspectives, and different stakeholders tend to instinctively adopt one of them. In this section, we seek to draw attention to issues in a multi-perspective manner. A similar approach has been used to consider interoperability and, although we have portrayed our idea of sharing as focussing on an inter-organisational relationship, and interoperability is often seen as being about IT-system level interactions, we embrace the view that these are inter-related perspectives. Indeed, the European Interoperability Framework\(^9\) (EIF) emphasises this unified view of interoperability and provides a useful structure to consider issues for data sharing (see Figure 1).

![Figure 1: Four Levels of Interoperability (from EIF v2, Figure 4-1)](image)

The following discussion of issues broadly follows the EIF, working from the bottom up, while aspects of the Political Context are touched-upon in section A 5 Roadmap.

At the end of each section of the discussion, a few synoptic problem statements are presented to capture the main points and to give an assessment of the proximity of the problem. The proximity of the problem is graded as follows, with numerical levels given to show rank order as for the aims.

1. Horizon: the issue will have impact but it could reasonably be addressed with low-level effort at present.

---

\(^9\) This is formally known as “Annex 2 to the Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions ‘Towards interoperability for European public services’”, or informally as EIF v2. It is available from [http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf](http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf)
2. **Approaching**: the issue will have impact in the mid-term and neglecting it could make matters worse in the future, for example by flawed technologies or practices coming into use. The issue deserves early attention.

3. **Blocker**: the issue is a blocker to meaningful progress.

### 3.1 Technical and semantic issues

The Description of Work of the LACE project identifies four core issues related to data sharing and interoperability. The first of these is the stimulus for this report on data sharing, while the remaining three issues emphasise aspects spanning the technical and semantic interoperability levels:

- Learning analytics and educational data mining research could make more progress in developing and testing algorithms and models if datasets could be more easily merged (both increased in scale and diversity of data).
- At present, analysts have to independently map between different data models. It would be more efficient for them if this work was done once and used many times.
- The absence of (data) interoperability means that there is wasted time and money spent to create and maintain computer code for multiple data exchanges. This makes it hard for suppliers to service localised or niche markets and hard for innovative products to become established. Few, if any, analytics-based applications will be viable – commercial or Open Source – except those provided by already-dominant suppliers of learning platform software.
- Existing specifications and standards have been neither motivated by, or designed for, learning analytics.

These are foundational issues for the data-sharing aims set out above, as well as system integration scenarios which do not fall within our scope for this report (Cooper, 2014a). From the point of view of practical application, the technical/semantic distinction is often not made by system developers, who usually deal with data structures and APIs as manifest entities. From the point of view of understanding the issues, however, and the possible actions to enable realisation of the aims we have outlined, it is useful to draw some distinctions. It is apparent, however that the technical/semantic distinction is somewhat arbitrary and that there are really various conceptual models from the technical representation of data at one end to facets of pedagogy at the other end. We draw the following coarse distinctions: low level semantics deal with basic representations, while domain-specific semantics would be appreciated by a teaching and learning practitioner, and in between these is the space of generalised meaning. These three levels are expanded on below, and comments made about the issues for learning analytics data sharing. The interoperability specifications referred to are further described in (Cooper, 2014b). It is also appropriate to consider the technical architecture here; it is not simply the representation and intended meaning of data elements which are necessary for practical interoperability.

### 3.2.1 Low Level

Low level semantics give meaning to the components of XML, CSV, JSON, relational database, etc structures. These are well defined and at a surface level pose no challenge to data sharing for learning analytics. Each, however, has its own form of expression, which colours the representation of data and so presents some practical difficulty.
3.2.2 Mid Level
Mid-level semantics are those in which the meanings of information elements are defined but only in relation to generalised contexts of use. For example, the Activity Streams\(^{10}\) specification defines a structure for making statements as part of social web activity which take the form Actor-Verb-Object (e.g. “Tore” “posted” “https://youtu.be/ufSPOJBB6og, Alyssa Wise Interview ‘Learning Analytics Silent Storm’”). ADL Experience API\(^{11}\) is similar, but introduces some domain-specific semantics, particularly from ADL SCORM\(^{12}\). Another example is JSON-stat\(^{13}\), which provides a means of communicating “data cubes”, while remaining mute about the meaning of the data attributes or the entities they refer to. There are both competing and complementary-but-overlapping approaches, hence some issues for data sharing. It is also the case that there is limited practical experience in sharing data between independently developed software, and differing views on the effectiveness of existing approaches. These issues do represent significant practical obstacles to data sharing as well as being a risk to anyone seeking to invest in the development of infrastructure and tools.

3.2.3 Domain Level
Domain-specific semantics are those in which we are concerned with, in our case, either concepts that are rooted in pedagogy or in analytics. The analytics domain - including traditional statistics and machine learning - is generally more easily-defined since mathematics and algorithmics are formal and objective in character. The data mining industry has developed PMML, the Predictive Modelling Markup Language, to allow for data transformations and other pre-processing steps, algorithm selection, and fitted parameters, etc., to be exchanged. There are several independent implementations of PMML and, in relation to learning analytics data sharing, we have found no evidence that it would be inadequate. Hence we conclude that interoperability in the analytics domain is not currently an obstacle in principle, although the practical reality is that PMML is not yet widely adopted for learning analytics (for an exception, see OAAI, section B 3.3 Open Academic Analytics Initiative (OAAI), and note that Jisc intends PMML to be an enabler for its open architecture, section B 2.6 Requiring an open architecture for learning analytics – the case of JISC plans for the UK).

The pedagogic domain is, however, a contrasting story because it contains great diversity. At present this diversity is poorly represented, and the quality of vocabularies that have been created is questionable; many have been created and used from a software-centric perspective with tacit assumptions about the teaching and learning domain\(^{14}\). Additionally, the true meaning of records of learning activities, hence the actual value of them, is not simply a matter of the objective description of what happened, no matter how rooted in the pedagogic domain it is. There may be numerous contextual factors, which may be difficult to express in objective form or may have a complex effect, which should be taken into account in deciding how to process the data, or what conclusions should be drawn.

\(^{10}\) [http://activitystrea.ms/](http://activitystrea.ms/)
\(^{11}\) [http://www.adlnet.gov/tla/experience-api](http://www.adlnet.gov/tla/experience-api)
\(^{13}\) [http://json-stat.org/](http://json-stat.org/)
\(^{14}\) This problem is recognised in parts of the industry, see for instance, the recent activities to form xAPI Communities of Practice, working on domain specific vocabularies for activity information - [http://www.adlnet.gov/tla/experience-api/xapi-cop-directory.html](http://www.adlnet.gov/tla/experience-api/xapi-cop-directory.html)
3.2.4 Technical Architecture

Sufficient common ground between technical architectures - we do not suppose there is a need for a single architecture, or that to pursue that aim would be anything more than a Quixotic task - is an obvious pre-requisite for technical interoperability, but effective technical architectures must embody both technical and non-technical requirements, including aspects which arise from questions of organisational interoperability and business models. When considering learning analytics, we must also face the possibility that the architectures which have evolved for technology enhanced learning may not be a good fit for learning analytics at scale. It is our contention that the current technical architectures for educational technology are not a good fit for learning analytic data sharing because they are not based on the requirements for analytics at scale, neither do they accommodate the non-technical requirements for privacy protection (which is expanded upon in section A 3.2 Organisational issues).

Summary technical and semantic issues

Issue TS1 - Incompatible technical implementations
Implementations of existing specifications, including choice of optional elements and vocabularies are not sufficiently-well aligned for data exchange between independently-implemented systems.

Issue TS2 - Inappropriate architectural assumptions
Many current early-stage approaches to learning analytics are unlikely to be viable in the long term because of being based either on historical architectures that are unfit, or because there is insufficient common ground in architectural approach to support multi-party interoperability.

Issue TS3 - Inadequate domain-level semantic models
Current domain-level information models are largely based on a software-centric, rather than pedagogy-centric viewpoint, and are generally poorly-defined.

Issue TS4 - Lack of adoption of existing specifications
There are candidate specifications for adoption that are largely being neglected, either through ignorance or lack of motivation. Whereas ADL Experience API is quite widely appreciated, appreciation of PMML, JSON-stat, etc. is low (Cooper, 2014b).

These are rated for proximity in Table 2.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 – Horizon</strong></td>
<td></td>
</tr>
<tr>
<td>TS4 – Lack of adoption of existing specifications</td>
<td>Whereas the adoption of existing specifications that are known to be usable is a plausible limitation on adopting learning analytics at scale, the impact on data sharing per se is questionable. General purpose data cube specifications such as JSON-stat are the exception, and these specifications would be useful for several Near-term and Broad Interest aims (level 3), i.e., Transparency in the success of learning analytics products and initiatives (aim 4), Cross-institutional strategy comparison (aim 5), and Research on the effect of education policy (aim 6). Of the existing educational technology interoperability specifications which exist, few could be recommended for adoption for learning analytics in advance of further work or adaptation, so these are not in scope for TS4.</td>
</tr>
<tr>
<td><strong>Level 2 – Approaching</strong></td>
<td></td>
</tr>
<tr>
<td>TS2 – Inappropriate architectural</td>
<td>This issue is most applicable to the operation of learning analytics at</td>
</tr>
</tbody>
</table>
specification of the problem, let alone the solution(s). The case of InBloom (section B 3.8 A counter-

| assumptions | scale in institutional, rather than academic research, contexts. In some ways, getting a response to TS2 as soon as possible is rather important since realising Aim 1 depends upon good-enough architectures, and this aim is being pursued by a number of organisations at present (level 4 aim). There is, however, an argument that the architectural requirements only emerge through practical experience and, so long as the TS2 is addressed as the field matures and significant investments are made with the view of this evolving picture, TS2 is not a blocker, but a potential cause of misadventure for which action should not be deferred. |
| TS3 – Inadequate domain-level semantic models | This issue applies across all the aims in section A 2 Exploring aims for data sharing because the interpretation, if not the processing, of learning analytics data must be based on an understanding of its significance in relation to the processes of teaching and learning. The extent to which this issue is a blocker or approaching depends on the radius of sharing. While data is shared close to its origin, it may be acceptable to apply coping strategies when using data expressed using inadequate conceptual models because the work-arounds and peculiarities are either known or discoverable by enquiry. Once you move far enough away from the data origin that its internal workings are difficult to determine, TS3 will risk bad learning analytics. Hence, although some progress can be made with things as they stand, TS3 requires some action now, although this will most likely be piecemeal. |
| Level 3 – Blocker | This issue applies across all the aims in section A 2 Exploring aims for data sharing, and although there has been some work on data exchange between independently-implemented systems (Downes et al., 2015), this work serves to make the point that TS1 is real and present. This issue must be resolved to move beyond insular data. |

Table 2: Rating of technical and semantic issues

3.2 Organisational issues

We found reason to mention it [privacy] in almost every chapter of this book. This surprised us. As the next chapter demonstrates we even found reasons to worry about privacy while looking at a box of cereal. (Robert Scoble and Shel Israel in “Age of Context”)

Our experience in the LACE project, in surveying the views of the learning analytics community, and in gathering case studies for this deliverable - see Part B - is very much in accord with the experience of Scoble and Israel, although we found that for learning analytics a broader set of questions pertaining to educational ethics should be bracketed with privacy matters. Consequently, in this section on Organisational Issues, we emphasise the topic of privacy and ethics, although we are conscious that this report has a very limited view on this broad vista. These are not new issues, neither are they newly-identified, but the fact that privacy and ethics are referred to so frequently in discourse on learning analytics indicates that they are critical issues. It seems likely that some stakeholders underestimate the complexity of the problem (section B 2.7 A Norwegian learning analytics vendor’s perspective on data sharing and interoperability), which could be characterised as being “messy” or “wicked” insomuch as it may not be realistic to reach agreement on the specification of the problem, let alone the solution(s). The case of InBloom (section B 3.8 A counter-
example: InBloom) illustrates one outcome when the problem is specified in ways that do not match the point of view of significant stakeholders.

In LACE, we have situated the issue of privacy alongside ethics, and introduced it in this section on organisational issues because we understand privacy as having its roots in the individual subject of the data. This is in contrast to some accounts of privacy which emphasise the legal constraints or which address IT security, typically the avoidance of security breaches. These may be necessary components of a response to the issue of privacy but are arguably insufficient by themselves because the focus on compliance marginalises the data subject (see section B 3.8 A counter-example: InBloom). Although, in this report, we introduce ethics and privacy in the section on Organisational Issues, important aspects are clearly in the Policy Context.

3.2.1 Ethics and privacy issues

Ethics and Privacy for Learning Analytics has emerged as a topic of strong interest for the communities LACE has engaged with. At the moment there are more questions than well defined positions or elaborated solutions. In Section PART B 1 LACE Workshop on Privacy and Ethics, we give a list of issues solicited through a number of workshops and activities engaging a wide range of learning analytics experts and stakeholders. For more questions and issues a taxonomy of ethical, legal and logistical issues has been developed by Niall Sclater (2015) after workshops organised by Jisc, Apereo and LACE. The taxonomy sorts the issues according to type (ethics, legalities or logistics), rank (importance attributed to the issues by experts in the field), responsibility (according to stakeholders involved, e.g., senior management, analytics committee, data scientist, educational researcher, IT department, or student). The resulting table of issues\(^{15}\), grouped by characteristics like ownership & control, consent, transparency, privacy, validity, access, action, adverse impact, and stewardship gives a blueprint for the stakeholder groups mentioned to start addressing the issues. It might, however, be noticed that a mere systematisation of questions, issues and concerns will bring us just a step forward, giving us a means to probe the quality of our proposed solutions. There is a need for a deep understanding of the multi-dimensional and cross-cutting nature of these issues to come up with solutions that are both technically efficient as well as ethically and pedagogically well founded. To achieve this goal we also need to look beyond education and recognise that other domains also face ethics and privacy issues coping with a more data-driven society.

The Jisc taxonomy was used as input to the development of a Code of Practice for learning analytics. The Open University (2014a, 2014b) has published a Policy on Ethical use of Student Data for Learning Analytics, which has been met with great interest from the field. In this policy The Open University ascertain that “learning analytics is an ethical practice that should align with core organisational principles, such as open entry to undergraduate level study”. In developing such policies on ethical use of student data one should observe some more generalised principles originating from discourse mainly taking place outside education.

Privacy as maintenance of Contextual Integrity

The debate on privacy tends to swing between two positions, one focusing on allowing individuals to control their personal information, the other limiting the number of persons gaining access to personal information. The Contextual Integrity perspective moves the debate beyond ‘control’ and

\(^{15}\) [http://analytics.jiscinvolve.org/wp/2015/03/03/a-taxonomy-of-ethical-legal-and-logistical-issues-of-learning-analytics-v1-0/](http://analytics.jiscinvolve.org/wp/2015/03/03/a-taxonomy-of-ethical-legal-and-logistical-issues-of-learning-analytics-v1-0/)
‘limitation’, promoting respect for context as a benchmark for privacy online (Nissenbaum, 2014). “When we find people reacting with surprise, annoyance, indignation, and protest that their privacy has been compromised, we will find that informational norms have been contravened, that contextual integrity has been violated” (Nissenbaum, 2014). This perspective emphasises context as social domain, and warn against giving primacy to context as technology system or platform; or context as business model or business practice; or context as sector or industry. SectionPART B 1 LACE Workshop on Privacy and Ethicsidentifies contextual integrity as an issue.

By applying context integrity as a social phenomenon the negotiable aspects of privacy is foregrounded. From this perspective, the institution may not have violated the informational norm if the roles of the actors involved, e.g. students, teachers, administrators, are acknowledged; the agreed information types are used; and the agreed data flow terms and conditions are followed. Seeing privacy from the perspective of maintenance of contextual integrity have many implications for design, both of organisational and technical solutions. Underlining the social dimension of privacy aligns with an understanding of learning as conversational activity, ref Laurillard’s conversational framework for the effective use of learning technologies (Laurillard, 2013)

**Learning and privacy as a risk-based activity**

Learning as part of human existence comes with a risk. From the health domain we know that we are willing to give up some privacy if we gain in terms of personal health. This does not mean that we should accept leakage of personal identifiable information. But we know there are risks involved, and with the current development of DNA-based information scaringly so. We know that we will not stop sharing our DNA-profile if our health is at risk, even if we know that it would be possible to derive a genotyptic barcode by matching facial heritable information by use of a public surveillance camera (Schadt, 2012). SectionPART B 1 LACE Workshop on Privacy and Ethicsidentifies the underpinning trade-off between benefits and risk. Observing privacy in a health context opens up an understanding of privacy as a risk management issue, which calls for more education and legislation. The same applies for education.

**Linking privacy to technical architecture**

Effective technical architectures must embody both technical and non-technical requirements, including aspects which arise from questions of organisational interoperability. Consequently, an architecture which is assumed but not based on an exploration of representative requirements will probably be unfit. It is our contention that the current technical architectures for educational technology are not a good fit for learning analytic data sharing.

An example may help to illustrate the point. A challenging requirement, not yet addressed, is the need for trust and identity management capabilities to be embedded into the learning analytics systems. In the case study of the Norwegian Connect service (section B 2.5 Building a cross-sector service platform – the case of Connect, Norway) learning analytics requirements were not accounted for, although the solution could probably have been shaped to meet some of the privacy and trust building requirements we have identified in this report. If and when the requirements addressing the issues related to access to data are fully recognised in technical design this might have wide ranging implications for the scope of learning analytics applications and how these solutions could be scaled up in learning, education and training.
Technical innovations (e.g., sensors, cloud computing, embedded processors, etc.) have had immense organisational impact in the field of analytics (see for example section B 2.2 PELARS – a European project investigating practice-based learning for an educational example, but the point being made here is about the pervasive nature of these innovations). However, organisational issues (e.g., lack of trust, and data protection) should also spur technical development. The modest technical development so far (specifications for expressing activity streams, activity ontologies, learning record store, etc.) has been based on a rather simplistic implementation model, which has excluded some of the psychological, ethical, and organisational richness of learning, education and training. When these soft issues are turned into hard requirements for technical development, we see that there is a need for redesign of some of the architectural preconditions for learning analytics solutions. As an example, when trust is recognised as a decisive issue for sharing data, features should be built into the tools from the very beginning which allow the users to be confident that their trust is being respected. For example, at any stage, a user should be able to check what data is being shared with whom and for how long. An architecture designed to sequester data, no matter how secure against inadvertent disclosure or penetration, will not build trust through this kind of transparency.

**Privacy-by-design**

When designing new technical and organisational solutions one should observe that “the principles of data protection by design and data protection by default” have recently been built into European and US policies, respectively through the General Data Protection Regulation (European Commission, 2013), and Recommendations for Business and Policy-makers from the US Federal Trade Commission (FTC, 2012). The Privacy by Design (PbD) framework encompasses IT systems, accountable business practices, and physical design and networked infrastructures; following seven foundational principles, among them Privacy as the Default Setting; Privacy Embedded into Design; Full functionality; End-to-End Security; Visibility and Transparency; and User-Centricity (Cavoukian, 2014).

**3.2.2 Other organisational issues**

**Lack of data and analytics knowledge among decision-makers**

Different kinds of organisations will have different decision-makers on matters of learning analytics, and in some cases, notably the school sector, these people may be somewhat detached from the data subjects. From strategy-formation down to the creation of contractual artefacts such as data processing/sharing agreements, a level of understanding of the role and handling of data that is required that is not typical of decision-makers in education and training (section B 2.8 A Norwegian publisher with adaptive learning product cooperating with a US company and B 2.9 School owners’ concerns – what is allowed according to the Privacy Protection Act? The case of giving advice to Norwegian schools), and by no means widely-appreciated in the business world. The potential consequences are rather diffuse across a spectrum, from the stifling of worthwhile initiatives at one end, to investment in initiatives that are doomed to failure because they fail to anticipate technical or sociological limitations at the other end. Issues arising from decision-maker knowledge deficit in relation to legal matters is expanded upon in section A 3.3 Legal issues.

---

16 [https://www.privacybydesign.ca/](https://www.privacybydesign.ca/)
Summary organisational issues

Issue O1 – Privacy
This issue is a manifold combination of: the widespread occurrence of under-developed conceptions of privacy among stakeholders; the inherent messiness (“wickedness”) of the problem; and the absence of proven approaches.

Issue O2 – Inadequate decision-maker knowledge
Decision-makers have incomplete or inaccurate mental models of the territory around learning analytics data sharing. See also Issue L1 (section A 3.3 Legal issues).

These are rated for proximity in Table 3.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 – Horizon</strong></td>
<td></td>
</tr>
<tr>
<td>O2 – Inadequate decision-maker knowledge</td>
<td>This issue has been placed at level 1 since, once the legislation-specific aspects are separated (section A 3.3 Legal issues), the residual impact on data sharing is difficult to pin down. The issue is, however, not to be dismissed and were this report considering learning analytics in general, the issue of decision-maker knowledge would be given higher priority.</td>
</tr>
<tr>
<td><strong>Level 2 – Approaching</strong></td>
<td></td>
</tr>
</tbody>
</table>
| O1 – Privacy | Although initiatives concerned with Aim 3 (A critical mass of data for learning science research) have working approaches to privacy protection, it is clear that many stakeholders appreciate the issue is something of a quagmire and that debacles such as InBloom reveal the risks of failing to address it adequately. The significance of O1 across the range of aims (section A 2 Exploring aims for data sharing) varies greatly but its role in relation to Aim 1 (More useful analysis through the combination of data from different sources) demands prompt action, and much of what might be aspired to is currently blocked. O1 has been classified as a level 2 issue because some progress is possible without change, although this could be considered highly risky as there may be an uprising of opposition from data subjects (especially learners).

It is anticipated that, in addition to varying significance, there will be a diversity of approaches between aims, and that appreciation of the issue and approaches to deal with it will evolve over time. |
| **Level 3 – Blocker** | |
| (none) | |

Table 3: Rating of organisational issues

3.3 Legal issues
The legal aspects of privacy-protection are apparent in the previous section, representing issues that are real and more likely to be appreciated by institutional or business management than some of the softer issues. In general, the issues arising from existing legislation are more easily attended to than the softer issues, and are well-understood by legal professionals. The problem is that what is possible within the law may not be well understood by other professionals. An over-cautious approach presents an obstacle (section B 2.9 School owners’ concerns – what is allowed according to the Privacy Protection Act? The case of giving advice to Norwegian schools). Indeed, the existence of legislation that deals with privacy and well-trodden paths to compliance leads some stakeholders to
equate compliance with privacy-protection; the relatively-lower awareness of the soft aspects presents an issue in its own right. This is not strictly a legal issue, rather it is a perceptual issue, but it is probably the most immediate legally-related challenge to overcome.

There are two further points we wish to make, which are maybe less pressing; international data sharing, or even within the EU, poses particular issues (see e.g. section B 2.8 A Norwegian publisher with adaptive learning product cooperating with a US company):

- Legal differences between the EU and the USA are a well-known issue in international commerce and the “Safe Harbor” process has been established to make it easy for US-based companies to comply with EU directives on the protection of personal data. While it is likely that a significant public procurement would verify compliance with the Safe Harbor Principles, it cannot be assumed that all services that are available for consumption in the EU are compliant.
- In spite of their being a common approach to data protection and privacy (e.g. the right to be forgotten) within the EU, member state legislation is far from uniform. This presents problems for providers of learning analytics services who wish to have a pan-European market, which potentially constrains market development and innovation within the EU as providers are likely to risk averse - bad publicity over a privacy-related issues could be terminal for a business - and to lack the resources to properly understand the variation.

A focus solely on current legislation, however, misses some important issues; it is plausible to assert that existing legislation has arisen from a historical conception of the problem and of technology. The resolution of anachronistic legislation can only occur through political process, so this issue could have been placed under the heading “Political Context”.

Summary legal issues

**Issue L1 – Lack of awareness of what is possible within the law**
While noting that compliance with the law does not equate to what should be done, it is apparent that some kinds of data sharing may be being ruled out either due to over-caution or due to lack of awareness of how to achieve the desired ends within the law. Issue O2 (section A 3.2 Organisational issues) could be expanded to include these legally-related matters, but we have kept L1 as covering those aspects which related immediately to legislation.

**Issue L2 – National differences**
National differences in data protection and privacy law within the EU impede access to an EU-wide market for learning analytics products. Legal differences with the US complicates procurement.

**Issue L3 – Current legislation is out of date**
Legislation trails behind the changes in the ways data is being used, and the evolution of the way citizens relate to technology change.

These are rated for proximity in Table 4.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 – Horizon</strong></td>
<td></td>
</tr>
<tr>
<td>L2 – National differences</td>
<td>The issue of national differences is certainly one of importance to</td>
</tr>
<tr>
<td></td>
<td>market development in general, and impinges particularly on Aim 1</td>
</tr>
</tbody>
</table>
Data Sharing Requirements And Roadmap

(More useful analysis through the combination of data from different sources). The natural difficulties that arise when attempting to use data originating in different education systems suggests that L2 could remain “on the horizon” except that the core of this issue is about the obstacles to using the same software components across borders (not the same instances of the software), and variety introduces cost. At present, however, we see this Aim 1 being pursued at national or sub-national level and while the understanding of learning analytics data sharing is maturing at that level, L2 is not yet close-to.

<table>
<thead>
<tr>
<th>Level 3 – Blocker</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3 – Current legislation is out of date</td>
</tr>
<tr>
<td>Legislation trailing technology change certainly has a large overall effect on data handling in general, but the extent to which L3 is impeding the development of data sharing for learning analytics across any of the aims in section A 2 Exploring aims for data sharing is low. While L1 remains as a target for action, it would be hard to claim L3 is anything more than a future obstacle.</td>
</tr>
</tbody>
</table>

**Table 4: Rating of legal issues**

<table>
<thead>
<tr>
<th>Level 2 – Approaching</th>
<th>(none)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 3 – Blocker</td>
<td></td>
</tr>
<tr>
<td>L1 – Lack of awareness of what is possible within the law</td>
<td>Although some data sharing activity can begin while this issue exists, it is a blocker in some sectors and for useful activity, particularly in relation to data sharing for other than research purposes, but potentially also for Aim 4a (Reproducible research).</td>
</tr>
</tbody>
</table>
4 Addressing issues

The issues which we previously identified have generally been expressed in fairly general terms such that we can be open-minded about how to address them. The principle is to keep options open since we are discussing the issue of data sharing in relation to several possible aims; in some senses this report reflects a search for “low hanging fruit” (and an identification of what might be out of reach). This section explores how the more pressing issues may be addressed in principle, and indicates some practical action which could make progress along that direction. Issues which are on the horizon are put to one side.

The following description of possible approaches to the issues identified in section A 3 Issues for data sharing are organised according to the same sub-headings as used in that section. Under each sub-heading, groups of 1 or more related action are presented under a title prefixed with an identifier starting “A-“ - e.g. A-TS1 - to be used to refer to that action-group. The numbering of these does not necessarily correlate with the identifiers used in the section on issues. A consideration of the relationship between action-groups, issues, and aims is left to section A 5 Roadmap. Each subsection concludes with an assessment of the feasibility of the action-groups, with a brief indication of significant variation depending on the aim (section A 2 Exploring aims for data sharing).

The feasibility levels are, in order of increasing feasibility:

1. **Challenging** – work is required to properly understand the issues in order to guide action and it may not yet be clear what is feasible. Action may be contentious or limited action would be possible without better problem definition, although this would probably involve an iterative process of: define, act, refine.

2. **Involved** – this would be likely to require some concerted effort, and probably several cycles of development, but the parameters for deciding what success looks like are largely uncontentious and known, or could be determined early in any project.

3. **Uncomplicated** – projects/initiatives could be established with a high likelihood of achieving useful outcomes in a single cycle of action.

4.1 Addressing technical and semantic issues

The account of the issues drew some distinctions between the problems at different levels of viewing the technical and semantic aspects of interoperability, and noted the increasing level of difficulty moving from low- to domain-level data definitions. The current situation is one in which the challenge is quite generalised; we have not detected any particular areas where stakeholders have identified a burning issue requiring targeted activity. This is believed to arise from the very limited extent to which shared data is being used; although ADL xAPI is, for example, being used quite widely, the analytics being undertaken is generally quite limited and is often in the context of a single system. The current emphasis on data capture rather than on use means that the issues are somewhat latent. Hence, we propose activities that are supportive of a process through which the more detailed issues are identified and addressed, rather than suggesting activities to target particular application areas or classes of data.

**A-TS1 – Shared open architectures and common frameworks**

The over-arching issue of increasing the adoption of interoperability would be aided by common reference-points to orient more localised activity (see e.g. section B 2.7 A Norwegian learning analytics vendor’s perspective on data sharing and interoperability). These enable the crossing of the
threshold at which the network-effect of benefits exceeds the localised cost. There is a balance to be struck between anticipating the architecture/framework with a thought leader (e.g. Jisc, section B 2.6 Requiring an open architecture for learning analytics – the case of JISC plans for the UK, or OAAI, section B 3.3 Open Academic Analytics Initiative (OAAI)) driving change vs. allowing it to emerge organically. The most-effective approach may well be for initiatives such as the Open Learning Analytics Network to work as the venue for drivers to calibrate their ideas with peers and to subject them to constructive criticism from experienced and knowledgeable stakeholders. Critically, though, change should be driven by public and private sector bodies working together, conceivably in a range of differing configurations, with the intent of developing systems people actually use.

A-TS2 – Code-bashes – mid-level practical interoperability
The issue of reliable exchange at the mid-level was identified as a practical barrier to higher levels of interoperability. This practical barrier is best addressed by practical action, but isolated from “line of business” or “production” implementations, and is often addressed by events variously called “code-bashes”\(^\text{17}\) or “plug-fests” which get people actually exchanging data between heterogeneous and independently-developed systems (Campbell & Cooper, 2013). In some of these events, the intention is for software developers to iteratively find and fix bugs in software or errors in interpretation of the interoperability specification, or ambiguities and errors in the specification itself. Although there has been some activity relevant to learning analytics so far - IMS has undertaken software development within its Caliper project and three Experience API Learning Record Store providers recently operated a data-exchange project (Downes et al., 2015) - there remains a lot of potential benefit from increasing this activity and in undertaking it in a variety of education/training contexts. Exploration in a variety of contexts is important because different contexts will exercise different aspects of specifications and different possible ways of interpreting their conceptual models.

A-TS3 – Practice-oriented pre-standardisation at the domain-level
Issues around the domain-level, contextual and pedagogical relevance, will require more work on defining structures, models, vocabularies through use in practice, as opposed to developer- or theory-driven definitions. This work should include an improved understanding of requirements, which would orient activity towards the more profitable areas and increase the long-term fitness for purpose. At present we have a combination of piecemeal activity in organisations such as IMS Global, Advanced Distributed Learning (ADL), and ISO/IEC JTC 1/SC36, and work from LACE to catalogue relevant specifications and standards (Cooper, 2014b), but these activities do not involve many people with domain-level expertise.

There is scope to involve learning analytics practitioner\(^\text{18}\) communities in practice-oriented pre-standardisation, defining things in parallel to learning-together. This kind of job-role is not yet widespread and these communities are only just emerging at present. Reasonable pre-requisites for addressing this issue are, therefore: facilitation of the emergence of these communities (and the identity of the learning analytics practitioner), and the coordination of pre-standardisation knowledge creation.

\(^{17}\) [http://blogs.cetis.org.uk/lmc/2012/12/05/codebashes/](http://blogs.cetis.org.uk/lmc/2012/12/05/codebashes/)

\(^{18}\) i.e. people who use learning analytics for institutional research or action research, rather than academic research.
Feasibility of actions

<table>
<thead>
<tr>
<th>Action</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 – Challenging</strong></td>
<td></td>
</tr>
<tr>
<td>A-TS3 – Practice-oriented pre-standardisation at the domain-level</td>
<td>Although the idea of practice-oriented pre-standardisation sounds quite feasible, our experience of standardisation and pre-standardisation activity suggests it is not an easy thing to do. Standards bodies have a poor track record of engaging with, or promoting, this kind of activity and research work is rarely set in the wider strategic perspective on interoperability. Software suppliers seek a return on investment which makes them cautious about engagement. The approach advocated in A-TS3, which attempts to navigate some of these structural issues with the de facto state of pre-standardisation, is to work from the bottom-up, but the details will require an exploratory approach.</td>
</tr>
<tr>
<td><strong>Level 2 – Involved</strong></td>
<td></td>
</tr>
<tr>
<td>A-TS1 – Shared open architectures and common frameworks</td>
<td>Action towards shared and open architectures and common frameworks has already begun and the experience so far has shown that this must be an iterative process. The current need is to sustain and expand activity, to feed lessons-learned and forward thinking into the knowledge-creation process. Success will be indicated by a growing buy-in to a number of stereotypes.</td>
</tr>
<tr>
<td>A-TS2 – Code-bashes – mid-level practical interoperability</td>
<td>Previous extensive use of this kind of action shows it is a practical way forward, but also one where improvement is incremental, both as technical complexity is advanced but also because of the link between practical interoperability and context of use (as specifications are put to new uses, more issues are revealed).</td>
</tr>
<tr>
<td><strong>Level 3 – Uncomplicated</strong></td>
<td></td>
</tr>
<tr>
<td>(none)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Rating of feasibility for technical and semantic issues

4.2 Addressing organisational issues

Data ownership, consent to the use of data, anonymisation and threat of re-identification, trust in systems and actors are only some of the organisational issues identified during LACE community exchange and analysis of the case studies in Part II. The issues around privacy suggest a greater variety in strategies than were considered for technical/semantic interoperability. Some of these might invite lateral thinking. For example, instead of selling or sharing data, organisations owning data sets could sell or share privacy-preserving data analytic services based on these data sets. “The question to be addressed then is: how would the business model around data change if privacy-preserving data analytic tools were available?” Thuraisingham (2014). We do not develop this line of exploration, noting it as a stimulus for brainstorming.

Addressing organisational issues, there are both “hard” and “soft approaches, bearing in mind the complex relationships between processes, policies (including rights and responsibilities), control, and technical advances when considering action to deal with these issues. The following list is ordered with the harder approaches at the start, concluding with an idea which is more like a meta-solution, or a unifying idea (risk-based approaches).
A-01 – Anonymisation and statistical disclosure control (SDC)
A normal response to the issue of privacy protection in shared data is to assert that it should be anonymised. The term “anonymisation” is often used to mean the removal of personally identifiable information (PII), with the understanding that PII is a set of pre-defined attributes, often defined in relation to legislation. Collected data, even if anonymised by removing identifiers such as names or other personally identifiable information, when linked with other data may lead to the re-identification of the individuals to which specific data items are related to. While the limitations of common approaches to anonymisation (sic) are known to people who have participated in discussions on privacy and ethics in learning analytics (section PART B 1 LACE Workshop on Privacy and Ethics), they are not raised as issues by the wider group of stakeholders.

Re-identification is known to be easy when unusual attributes or particular events are captured in the data and when other information sources contain the same facts. For example, given some data points referred to a mature female student on a mechanical engineering course at a given university, it would probably be possible to associate that data with an individual from public-access information. One of the challenges is to assess the re-identification risk without knowing in advance what other data is available to a snooper (Martin et al., 2007). Timestamped geolocation data is also very easily used to identify the individual to whom it relates (De Montjoye et al., 2013) and by extension could be used to identify familial or social groups with intrusive levels of probability. Sometimes less-clear, but still potentially undesirable inferences may be drawn from data. For example, postal codes may be strongly associated with ethnicity, which may be an intrusion by itself or the means to increase the probability of re-identification. Textual data is another cause of failure in attempts to anonymise data, especially as it arises in online forums (Teutsch et al., 2009), in which people may be referred to by name, posters will commonly sign their own name, email footers may disclose other contact details, or posting and replying by email may leave email addresses in the text.

This does not mean that we necessarily have to give up privacy in order to use big data for good societal purposes. It does mean that, although the term “anonymisation” is commonly used, it is not accurate; whether the data is anonymised or not is a post-hoc discovery, not a guaranteed consequence of the process. The term “statistical disclosure control”, which is used by the UK Government Office for National Statistics (ONS), among others, may be a better one to use for the process often called “anonymisation”, but it also includes some ideas which extend scope a little, for example synthetic data (see later). The ONS describes SDC as “Statistical disclosure control involves modifying data or outputs so that the risk of identifying individuals is reduced to an acceptable level.” It is a topic for which there are substantial international efforts with, for example well-supported workshops organised by Eurostat (Eurostat, 2009).

Although anonymisation (sic) and statistical disclosure control is, in a general sense, a means of addressing part of the issue of privacy, it is useful to be more specific, especially to consider actions to address the issue in the context of learning analytics data sharing. An alternative formulation of this statement is to state that we should consider the issues pertaining to the current approaches to anonymisation for learning analytics. The suggested actions and related subsidiary issues are:

- Spread knowledge of the limitations of simply removing PII, to avoid flawed approaches being widely used, with consequent risk of disclosure or reputational damage.
• Undertake research to quantify the risk of disclosure in real learning analytics data-sets.
• Undertake research on the utility of existing SDC methods from statistics agencies for learning analytics datasets and data sharing aims, including the possible benefits of creating synthetic data (i.e. data which replicates features of the original dataset without containing any of the original records). This work should include the development of protocols (widely accepted repeatable practical methods) appropriate for different situations, and maybe practical software tools. There is a considerable quantity of research and practical use of SDC for household survey data but a remaining need to explore how these methods relate to the data, processes, aims, and constraints of learning analytics practice and research.
• Undertake further research on the de-identification of user-generated text, and the development of practical tools appropriate for the TEL context.
• Stimulate activity to promote the absorption of advanced methods into learning analytics. For instance, recent advances in cryptography are making it possible to perform analytics on encrypted data, although it is expensive, on several axes.

A-O2 – Analytics models as shared data
We have largely been considering either original person-level data (“microdata”) or aggregate-data statistics (“tabular data”), which has a lower, but non-zero, risk of disclosure; however, an alternative is to consider sharing the analytical models. These models, which might be predictive, descriptive, or explanatory of generative processes, etc., are not usually seen as being data and they are rarely shared except in research papers, and often then the information is incomplete. Sharing a fully-elaborated model (e.g. a predictive model, sufficient to drive a prediction engine), for example using a standard format such as PMML, could address some data sharing aims while allowing disclosure risk to be contained. A useful set of actions would be to investigate the practicality of this approach in the context of particular data sharing aims.

A-O3 – Remote access analytics
We know anonymisation is not a panacea, and even with advances in statistical disclosure control and a more balanced appreciation of risk, and a willingness to embrace risk-based approaches (see below), there may be occasions when it may be possible to circumvent the problem in other ways. For some tasks we could establish remote access systems and policies that allow analytics on data without release of the data. The remote access system would allow on a pre-defined set of pre-processing, statistical, or machine learning methods to be applied and impose access control and SDC. The idea of remote access analytics could be applied to several of the aims for data sharing (section A 2 Exploring aims for data sharing), and it is to be expected that quite different realisations of the idea would apply; “remote access analytics” is not a single system.

Existing cloud-based analytics services, for example Zementis, demonstrate that the central idea of a remote analytics service is technically-feasible and interoperability standards exist for this purpose (PMML). Further work is required to establish whether controls could be imposed, for example by checking the PMML, which meet stakeholder requirements for acceptable risk, but the idea of a secure location where data-contributors can only see a segment of the data but undertake analytics across the whole set is credible. The OpenPDS/SafeAnswers\(^\text{19}\) system, which is available as Open Source Software, is an MIT Human Dynamics Group project with international partners that has

\(^{19}\) http://openpds.media.mit.edu/
conceptualised the idea of remote access analytics as a personal service (De Montjoye et al., 2014). The service isolates mobile/web apps from directly accessing the low level data about location, for example, and acts as an intermediary, giving degraded, aggregated, or otherwise sanitised information which is still useful for the app. The SafeAnswers system has been built with ubiquitous smartphone/tablet use in mind, but it provides a good step-up for imagining how learning analytics at an organisational level could co-exist with learning activity data as a personal database. The contrast between the app-friendly SafeAnswers and the full-blown analytics of Zementis illustrates the point that remote access analytics is a concept, not a single system.

A-O4 – Trusted data analysts
An alternative approach to Remote Access Analytics is to re-cast “remote” in human organisational terms. For some tasks we could limit disclosure working through a trusted agent who receives data from multiple sources and process it, returning the results to each stakeholder. These results could have been arrived at by the use of the full dataset. This approach is used by the PAR Framework20.

A-O5 – User-managed access
More research is needed in data confidentiality techniques and privacy-preserving techniques which include the data subject in a more sophisticated way than “opt in or out” (sectionPART B 1 LACE Workshop on Privacy and Ethicsindicates stakeholder concern with the utility of this binary approach) and which engender trust, for example through transparency. For the educational community this implies a need for defining the privacy needs for the different stakeholders taking part in learning, education and training. These should shape the technical system architecture, user experience, and surrounding policies. One emerging approach is to give control over how the data is used to the data subject, which is the focus of the Kantara Initiative Work Group on User Managed Access21 (UMA). The UMA is underpinned by an architecture which assumes data sources, storage, processing, and control may be distributed, and would provide a technical starting-point for a consideration of UMA for learning-related data but the realisation of a practical system for education or training must account for the capabilities, habits, attitudes, etc. of the users, and that will vary dramatically between primary and secondary school, university, workplace, and informal learning. To be meaningful, consent must be informed, but it will also be important to avoid consent fatigue or unnecessary denial of access to services in the absence of consent. Furthermore, a practical realisation of UMA implies a range of policy-forming activity. The European Cookie Law is not, to judge by widespread criticism, a model to replicate.

A-O6 – Common codes of practice and standardised data agreements
Codes of practice would introduce efficiency by reducing the quantity of ad hoc reasoning about appropriate courses of action. The creation and adoption of common codes of practice (see e.g. section B 2.6 Requiring an open architecture for learning analytics – the case of JISC plans for the UKfor an initiative aimed at a common approach, while the case study B 2.3 LEA’S BOX – a European project creating a Learning Analytics Toolboxshows a project-level approach) would not only be desirable in terms of the saved effort from avoiding multiple very-similar code-development activities, but would help to promote alignment between organisations, i.e. it would enhance organisational interoperability. This would reduce risk arising from mis-match, for example in how

20 [http://www.parframework.org](http://www.parframework.org)
21 [https://kantarainitiative.org/confluence/display/uma/Home](https://kantarainitiative.org/confluence/display/uma/Home)
issues such as transparency, accountability and control are realised (see section PART B 1 LACE Workshop on Privacy and Ethics), and would enable sharing.

The Asilomar Convention for Learning Research in Higher Education lists represents a research-oriented view, and although the Convention is not explicit about the concept of organisational interoperability (by any name), it is de facto an attempt to increase convergence on common principles. The existence of functioning, established, respected, trusted, and resourced mechanisms of approval for research on ethical grounds gives us a straight-forward baseline, from which to explore the regulation of data sharing other than for formal academic research. Existing initiatives to open-up datasets are very-much aligned to formal academic research - for example VPOL (section B 2.5 Building a cross-sector service platform – the case of Connect, Norway) states: “[it] is committed to enabling scientific inquiry through the provision of access to data from digitally mediated instruction to qualified researchers at Stanford University and worldwide” and several of the case studies in Part II illustrate the research data perspective (e.g. B 2.2 PELARS – a European project investigating practice-based learning, B 2.3 LEA’S BOX – a European project creating a Learning Analytics Toolbox).

To move beyond the working model for human subjects research, and to adapt it for operational learning analytics will require some developmental and piloting work to accommodate a different ethical milieu, power structure, etc. Such a system could, for example, function as a regulatory counterpart to User Managed Access.

Standardised data processing/sharing agreements provide a complementary tool with similar benefits (see, for example section B 2.5 Building a cross-sector service platform – the case of Connect, Norway). Within this action-group, we might expect a combination of actions, which develop approaches to the regulation of data sharing and use from the research ethics baseline; take a wider view of codes of practice with influence on data sharing processes and practice; and standardise contractual agreements with third parties.

A-O7 – Develop understanding and consensus around risk-based approaches to privacy protection

Education is not the only sector facing privacy protection issues in a more data-driven society. There is a lot to be learnt from health, environment, city planning, and even cyber security. From the discourse on security and privacy we see thoughts about a “multi-objective optimisation framework for data privacy” (Thuraisingham, 2014). There is no one size fits all solution for data privacy.

“Instead, multiple dimensions need to be tailored for different application domains to achieve practical solutions” (Thuraisingham, 2014). What is the ideal mix between privacy risks, costs and utility of privacy-enhancing techniques in learning? The process of answering this question will require a careful balance of understanding stakeholder sentiment and values, and a thorough understanding of how socio-technical systems work. As stated, the question could be applied to many areas of human activity where the increasing capture and use of data is occurring, and there is prior and ongoing research to understand it. This should be taken into account, but the need remains: to develop an understanding and consensus that is contextualised to learning analytics and the differing values that apply for application in schools, universities, the workplace, or elsewhere.

---

22 http://asilomar-highered.info/
Feasibility of actions

<table>
<thead>
<tr>
<th>Action</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 – Challenging</strong></td>
<td></td>
</tr>
<tr>
<td>A-O3 – Remote access analytics</td>
<td>Remote access analytics, while having some aspects that are demonstrably feasible (see the description of A-O3), has been classified as “challenging” because viable approaches would require both buy-in to an unfamiliar approach from stakeholders and some process for managing cost (although cloud computing dissolves some cost issues). It would be possible to tighten scope, imposing limits on what “remote access” would entail, to increase feasibility to level 2. As a general concept, though, we don’t yet know what remote access analytics would actually look like.</td>
</tr>
<tr>
<td>A-O5 – User-managed access</td>
<td>User-managed access is an appealing concept which superficially appears to make some challenging issues disappear by transferring control. Beneath this skin is a mess of complexity around user experience, policy, system architecture, analytical validity, etc. To understand this mess, and to discover more of what UMA means, would require some exploratory piloting, undertaken in a way which allows the complexity to surface, rather than making assumptions and drilling deep into only one aspect.</td>
</tr>
<tr>
<td><strong>Level 2 – Involved</strong></td>
<td></td>
</tr>
<tr>
<td>A-O1 – Anonymisation and statistical disclosure control (SDC)</td>
<td>The substantial base of existing statistical and theoretical work that already exists outside the field of learning analytics indicates that a process whereby they are adopted and adapted into learning analytics by research and innovation activities would be quite feasible, although a multi-stranded and staged process.</td>
</tr>
<tr>
<td>A-O7 – Develop understanding and consensus around risk-based approaches to privacy protection</td>
<td>Developing understanding and consensus can build upon existing work in health and by national statistics agencies (see A-O1, which focuses on some of the technical counterparts for evaluating risk) but it will require a change of mind-set for “knowledge-of” to translate to “understanding and consensus” sufficient to lead to action.</td>
</tr>
<tr>
<td><strong>Level 3 – Uncomplicated</strong></td>
<td></td>
</tr>
<tr>
<td>A-O2 – Analytics models as shared data</td>
<td>Noting the scope of action as being about the practicality of sharing fully-articulated models (i.e. assuming a situation in which model sharing would be useful), A-O2 is quite feasible, as shown by OAAI for its application.</td>
</tr>
<tr>
<td>A-O4 – Trusted data analysts</td>
<td>The fact that the PAR Framework exists shows that A-O4 is feasible and it is arguable that localising the approach would be relatively uncomplicated; although PAR took several years to mature, this can be ascribed to discovering facts which generalise well.</td>
</tr>
<tr>
<td>A-O6 – Common codes of practice and standardised data agreements</td>
<td>While codes of practice and data sharing agreements will undoubtedly evolve as the field matures, the first hurdle to usable results is relatively uncomplicated and progress in this direction is already being made on some fronts.</td>
</tr>
</tbody>
</table>

Table 6: Rating of feasibility of organisational issues

4.3 Addressing legal issues

The legal issues related to transferring data between legal entities are a well-known issue, and the situation is complicated by legal differences between countries. Big Data adds complexity because it enables non-obvious facts to be inferred, including privacy-intruding aspects, potentially by linking to separate data sets; Big Data makes it less easy to see what the data-set can be used for. Often the value of data lies in its potential future uses, and this is not a situation that is accounted for in European data protection regulations.
In the section on issues, we asserted that a restrictive interpretation of current law in which numerous realisations of learning analytics are deemed “illegal” is not helpful for education and slows progress (section B 2.8 A Norwegian publisher with adaptive learning product cooperating with a US company). Resistance that is not grounded in the real data protection and privacy issues related to these new educational technologies is an immediate problem. In addition to raising awareness of risk-based approaches (see above), an achievable response is to take action to increase the level of informed decision-making.

**A-L1 – Raise awareness of what is possible within the law**

In many countries, the data protection authorities are being active working together with developers to guide a privacy-by-design approach to new solutions (section B 2.8 A Norwegian publisher with adaptive learning product cooperating with a US company). Such cooperation should be encouraged, as well as a more open debate on data protection and privacy issues related to learning analytics (section B 2.9 School owners’ concerns – what is allowed according to the Privacy Protection Act? The case of giving advice to Norwegian schools). Organisations that serve the national school, college, or university sectors should take a lead as intermediaries to avoid unrealistic demands on the data protection authorities. Exemplars of what is possible within the law, specific to national legislation and principles, should be created and disseminated.

**Feasibility of actions**

<table>
<thead>
<tr>
<th>Action</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level 1 – Challenging</strong></td>
<td>(none)</td>
</tr>
<tr>
<td><strong>Level 2 – Involved</strong></td>
<td>A-L1 – Raise awareness of what is possible within the law There are no intrinsic difficulties with this action-group; it is clear what success should look like and the steps would require widely-held skills. There are some complications arising from the limited capacity of data protection authorities but the main factor suggesting A-L1 should be classified as “involved” is interest and motivation. It is believed that the first step would be to get the necessary stakeholders to a position where they felt that there is something to be aware of, to change attitude from avoiding prosecution to discovering possibilities.</td>
</tr>
<tr>
<td><strong>Level 3 – Uncomplicated</strong></td>
<td>(none)</td>
</tr>
</tbody>
</table>

Table 7: Rating of feasibility of legal issues
5 Roadmap

In general, there is a need for more research and community exchange to frame, sort and prioritise the issues of data sharing for learning analytics. Admittedly, an optimistic research community this year put “Scaling up: Big Data to Big Impact” at the theme for Learning Analytics and Knowledge Conference 2015; however, in order to scale up a better understanding of legal, organisational, and technical and semantic aspects of data sharing must be established. Sections A 1 to A 4 have presented the authors’ analysis of the problem space and the solution space, and Part B complements this with a range of case studies and examples which others can use to understand the situation from information closer to the source.

In this section, we move towards an outline roadmap with our current state of understanding. What we present is really an overlay of related roadmaps, in which there are plausibly steps in common between different intended aims. It is something from which different stakeholders could extract components to match their aims as the basis for the formation of coordinated plans of action. To advance towards such a coordinated plan would require considerable effort to gather a variety of stakeholders together - a different set for many of the aims - and to jointly plan. This stage of strategy formation is not in scope in the current work, which sets out to lay the foundations for such planning by analysing the domain of aims, issues, and relevant practical actions.

5.1 Business and political context

In the introduction to section A 2 Exploring aims for data sharing, we introduced the European Interoperability Framework (EIF) as scaffolding for the discussion of issues and noted the way it placed the Political Context as a backdrop. The EIF is, however, motivated by a consideration of interoperability in public services, whereas our view of data sharing for learning analytics encompasses a wider spectrum: teaching and learning in public or private educational establishments, or in the context of employment; private sector technology providers; and research activity. For this report, therefore, we imagine a “Business Context” counterpart to “Political Context”; the essential character of these two is very similar - “cooperating partners with compatible visions, aligned priorities, and focussed objectives” - but the Business Context should also consider viable business models and matters of public or private sector innovation investment.

Since these are contextual factors, they often do not easily resolve to straightforward statements of issues or isolated actions yet they are often influential. The descriptions of the issues and actions to address them make a number of references to the need to account for contextual factors, and the purpose of this current section is to capture some of the broad themes with a business/political flavour, which may have a more diffuse relationship to action, but which should still influence the way roadmaps are realised.

5.1.1 Level playing field

One issue, which was raised in the Norwegian Vendor case study, but which is also a guiding principle in the Jisc case study is the idea of “levelling the playing ground”. This might be by standardisation means, giving access to user data, etc. to all in the market, making it impossible for some companies to use lock-in strategies or to use existing market penetration in one product category as a means to leverage sales in another by putting in place protectionist data access arrangements. Where actions are sponsored by the public sector, the aim of levelling the playing field in the interest of an efficient market is important. This may come down to putting existing
policies into practice, rather than developing new policy; there are many efforts to manage the market in this way in Europe, while leaving the competitive principles in place.

5.1.2 Collaboration over capital investment
The particular issue with capital investment arises from the current deficiency in available software relative to what is needed to address some of the issues in section A 3 Issues for data sharing. In the age of cloud computing, the cost of infrastructure can largely be transferred to a recurrent cost but the process of developing fit-for-purpose software, where “fit” relates to user experience, policy, and practice, represents a cost barrier to achieving some of the outlined aims for data sharing.

The path of least action might take various forms, depending on the complexity of the requirements, or the novelty of the solution etc., but cooperation on software, for example using Open Source Software licences and open-working principles, could partially address the capital investment problem. The opportunity here is not so much focussed on the licence, but on the opportunities for collaboration and the processes which it enables. Both private and public sector organisations could collaborate on the same software; public sector bodies might take a long-term view and be interested in implementation of the UMA specification, for example, while private sector interests might identify this as being an obstacle rather than a business opportunity, something that will become a utility rather than a differentiating business proposition. These collaborations would not have to be established from scratch; initiatives such as Kantara, the Apereo Foundation, or the R Project for Statistical Computing should be preferred as a route to greater productivity and sustainability and examples such as Unicon, which is a private company developing Open Source Software through a series of commercial contracts, demonstrates there is variety.

There is scope for more businesses to re-think their business models to make greater use of Open Source Software as something they would both contribute to and take advantage of. Procurement practices in the public and private sector are also often not Open Source friendly.

5.1.3 Sustainability of public sector initiatives
In the public sector, we need to find better ways of stimulating change than traditional project funding. Reducing the risk of abandonment or closure has no simple answer, but is not a hopeless case. One strategy is to build-out from existing business-critical services. The Feide system for federated identity management and access control in Norway (section B 2.5 Building a cross-sector service platform – the case of Connect, Norway) is an example of such an established system, which is further developed into a cross-service platform; and the Estonian Education Cloud (section B 2.1 Designing a national learning analytics system – the case of Estonia) could act as a similar stable core (we have not assessed the long term viability of this initiative). This strategy may apply for pieces in the learning analytics architecture jigsaw, so long as the requirements of different usage scenarios are properly aligned. For situations where there is no conceptually-adjacent existing service and something new must be instantiated, a reasonable strategy may be to seek distributed or collaborative approaches with built-in resilience coming from no single point of funding. This could take a form such as the PAR Framework (section B 3.2 Predictive Analytics Reporting Framework (PAR)) or may be where the Jisc approach will lead to (section B 2.6 Requiring an open architecture

23 https://kantarainitiative.org/
24 The Apereo Foundation is concerned with academic software. https://www.apereo.org/
25 http://www.r-project.org/
for learning analytics – the case of JISC plans for the UK). For research collaborations, a wholly distributed approach might work, with no central entity, per-organisation costs that are marginal and a peer-to-peer data-mirroring approach. These may require pump-priming for some software development and the establishment of common workflow. An alternative strategy would be to look at the emerging business models around Open Access scholarship generally and to either adapt these to data sharing, or align data sharing to current Open Access provision.

5.1.4 Privacy protection and policy – when legislation falls short
Legislation is, of course, not necessary for change: policy can be formed and implemented at many levels of governance where the outcome is not law. Instruments such as codes of practice and spreading knowledge of ethical practice have a role to play. These kinds of instruments were mentioned in section A 4.2 Addressing organisational issues, as codes of practice which focus on learning analytics, learning science research, etc. and where ethics was implicitly bound to the cultural norms of education and training. It is also possible to draw back, to take in a wider political scope.

In relation to the anachronism of current legislation (section A 3.3 Legal issues), a parallel with health-care data may be drawn. Erik E. Schadt, an American genetics and genomic science researcher said: “I believe education and legislation aimed less at protecting privacy and more at preventing discrimination will be key” (Schadt, 2012). He is advocating stricter and broader anti-discrimination regulations as a condition for our societies to respect individual rights while benefiting from the tremendous potential of big data more openly shared in the life sciences and medicine. Such a change in approach could have extensive ramifications for the use of education and training data, not just for learning analytics. Learning analytics data sharing would be merely a footnote in the wider scheme but thoughtful analysis, research, and spreading knowledge about the avoidance of unintentional discrimination would enrich it; discrimination may hide in the analytical methods or in the actions arising from the results of those analyses.

5.2 Concerning actors
To a limited extent, progress can be made through independent work in different establishments; there is a need for educational establishments to address their internal policies and procedures. Cooperation and sharing can achieve some efficiencies in making progress but that often requires coordination and facilitation, and there are many issues for which concerted collaboration as a research or innovation action is necessary.

We see a role for national bodies, generally specific to stages of education, as coordinators and facilitators of activity against those aspects which connect most closely with institutional activity, as opposed to research. This idea is reflected in the case study in section B 2.7 A Norwegian learning analytics vendor’s perspective on data sharing and interoperability. National bodies with an education focus would be well-placed to cooperate with, or consult, bodies which are external to the education system, for example data protection authorities. At the school stage, it is likely sector bodies would take the lead to produce replicable and adoptable products. At the university stage, a lighter-touch facilitative role would generally be appropriate. Examples of this kind of body in the Netherlands are Kennisnet and SURF, relating to schools and universities respectively.
European Schoolnet has engaged in LACE activities (LA was a theme in the 2014 Eminent conference). The network of 31 European Ministries of Education has a role to play as an upper-level coordinating agent, to realise potential for cross-border synergies.

It is more difficult to point to a coordinating agent of similar nature for European universities. UNIS, the European University Information Systems organisations put LA on the agenda for their EUNIS 2014 conference, and there might be scope for the organisation to take more coordinated action. The operators of existing initiatives involving data sharing may be significant actors, bearing in mind that it is unlikely that replication would be appropriate, given differences in aim and cultural or business context in Europe. The two most promising are the PSLC DataShop\(^\text{26}\) (section B 3.1 Pittsburgh Science of Learning (PSLC) DataShop) and PAR Framework\(^\text{27}\) (section B 3.2 Predictive Analytics Reporting Framework (PAR)).

For those aspects which relate most closely with research, as opposed to institutional concerns, EATEL is a potential actor, specifically via its “SIG dataTEL – Data-driven Research and Learning Analytics”\(^\text{28}\). SoLAR, the Society for Learning Analytics Research\(^\text{29}\), and the IEDMS, the International Educational Data Mining Society\(^\text{30}\), provide a complementary venue. SoLAR, in particular, is seeking to bridge the gap from research to practice. Furthermore, its Journal of Learning Analytics\(^\text{31}\) is an open access publication that is well placed to address matters of reproducible research. Both organisations could usefully encourage and stimulate research activity on learning analytics data sharing. SoLAR was also the venue for launching the concept of an Open Learning Analytics Platform, which was the basis from which an informal network was initialised in cooperation with the Apereo Foundation. This network, the Open Learning Analytics Network\(^\text{32}\), advanced a number of Principles at its first meeting, including transparency and shared data. The LACE project worked with the Apereo Learning Analytics Initiative to host a second network event, in Europe, and, although the Network remains informal, it provides a useful forum, given its emphasis on “open”.

There is, however, a limit to the impact that these coordinating entities and initiatives can have in the absence of substantive research and innovation activity. It is clear, therefore, that progress towards data sharing to work towards the aims outlined in the introductory section of this report, and the benefits which these could accrue, will require resourcing from national research agencies and the European Commission Horizon 2020 Programme.

Finally, it is worth noting that significant work in Europe take place in countries where English is not the universal working language, and thus often does not achieve the impact it deserves, relative to work undertaken in the English-speaking world. The French MULCE (Multimodal Contextualised Learner Corpus - see section B 3.4 Multimodal contextualised Learner Corpus Exchange (MULCE)) is a case in point. There is, therefore, an overarching assumption that the roadmaps should address this structural issue through continued action by the European Commission.

\(^{26}\) \url{http://www.pslcdatashop.org} \\
\(^{27}\) \url{http://www.parframework.org} \\
\(^{28}\) \url{http://ea-tel.eu/sig-datatel/} \\
\(^{29}\) \url{http://solarresearch.org/} \\
\(^{30}\) \url{http://www.educationaldatamining.org/} \\
\(^{31}\) \url{http://learning-analytics.info/} \\
\(^{32}\) \url{https://www.apereo.org/content/learning-analytics-initiative}
5.3 Suggested steps to take

Having first considered various aims for data sharing and organised them according to our assessment of importance to the learning analytics community, and then considered issues and their proximity as obstacles to working towards those aims, various practical actions were suggested to deal with the most important issues, along with an assessment of the feasibility of each. In addition to the contextual factors outlined at the start of section A 5 Roadmap, these constitute the raw material for the LACE Data Sharing Roadmap, which is shown in a graphical form in Figure 2. The rows align with the EIF-inspired subsection headings used in sections A 3 Issues for data sharing and A 4 Addressing issues and the columns give a broad indication of a time-line. The green-tinted blocks indicate actions that were classified as being “uncomplicated”, orange blocks indicate “involved” actions”, and pink blocks were those identified as being “challenging”. Contextual factors and other points made in A 5 Roadmap are summarised in round-ended boxes with a blue background, outside the Roadmap but within the surrounding box marked “Contextual”.

The actions have been placed in time according to need. If they relate to blocker issues or the more important aims, they being in the “Now” column, otherwise they start in the “Soon” column. No actions start “Later” because actions were only suggested in section A 4 Addressing issues for issues deemed to be on the horizon. Action A-O5 (User Managed Access) is arguably not necessary to overcome as the most immediate privacy issues but, given the likely time to maturity, we feel an early start would be desirable.
### Figure 2: The LACE Data Sharing Roadmap, first version

<table>
<thead>
<tr>
<th>Category</th>
<th>Now</th>
<th>Soon</th>
<th>Later</th>
<th>Reasoning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical &amp; Semantic</td>
<td></td>
<td>A-TS3 - Practice-oriented Pre-standarisation at the Domain-level</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A-TS1 - Shared Open Architectures and Common Frameworks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A-TS2 - Code-bashes - Mid-level Practical Interoperability</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A-O3 - Remote Access Analytics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A-O5 - User Managed Access</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organisational</td>
<td>A-O1 - Anonymisation and Statistical Disclosure Control (SDC)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A-O7 - Develop Understanding and Consensus Around Risk-based Approaches to Privacy Protection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A-O2 - Analytics Models as Shared Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A-O4 - Trusted Data Analysts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Legal</td>
<td>A-O6 - Common Codes of Practice and Standardised Data Agreements</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A-L1 - Raise Awareness of What is Possible Within the Law</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Contextual:**
  - Levelling the Playing Field
  - Sustainability of Public Sector Initiatives
  - Collaboration Over Capital Investment
  - Privacy Protection and Policy
  - Existing Initiatives are a Resource (avoid re-invention)
  - Ensuring non-English Speaking Countries are Not Marginalised
  - Involvement of Coordinating and Multiplier Actors
  - Cross-border products and services face privacy law differences
# 6 Conclusion

The arc of the enquiry for which this report is the output, travelling through an outline of the variety of aims for data sharing in learning analytics into the issues and some of the potential means of addressing those issues has revealed considerable variety across all of these aspects.

The variety of aims for data sharing (section A 2 Exploring aims for data sharing) makes clear that “data sharing for learning analytics” is not a single idea and, in retrospect, the task set out in the LACE Description of Work to “develop a prioritised roadmap for development of shared data repositories for learning analytics and learning science research” could have been expanded into an entire project. Indeed, the idea of “a roadmap” became questionable as the range of possible aims for data sharing was explored because the balance of emphasis between the issues varies. Hence, we conclude that, as groups of stakeholders marshal interest around certain aims, each should expect to develop their own roadmap. Our singular roadmap represents a high level synthesis of actions which are likely to be of interest to the most stakeholders.

In respect of the issues, the variety is revealed in the complex socio-technical matrix; superficially technology-related aspects quickly mingle with the non-technical and non-technical issues reveal causal aspects in the technology domain. For example, the fitness-for-purpose of existing interoperability specifications was challenged according to the lack of educationally-motivated modelling, and the roots of some of the privacy concerns arise from what has become technically possible with machine learning.

When it comes to the possible ways of addressing the issues, we again find variety. In some cases there appear to be technical solutions to problems - both in terms of ICT and statistics - but we also identified actions which rely on spreading knowledge and changing attitudes, for example to embrace risk-based approaches. At the heart of the more challenging technical actions, such as the development of suitable interoperability specifications or User Managed Access software, our deconstruction of the action required, linking back to the characterisation of the issue it addresses, shows that participative and iterative approaches are necessary. Approaches to privacy must involve the data subjects. Activities intended to model learning analytics information must have foundations in practice if the information is to be shared with meaning.

Although many aspects of data sharing are rather messy, it is also the case that there are a number of existing data sharing initiatives that have sustained themselves beyond the initial project stage (sectionPart B 3 Existing examples of data sharing for learning analytics). The controlled environment of research is evidently the most amenable to making progress, while at the other end of the spectrum, InBloom shows how an opaque manifold of multiple data sources and sinks can generate huge concern.
Part B
PART B 1 LACE Workshop on Privacy and Ethics

From the very beginning of the LACE project, when engaging with the different LA communities and starting to explore the focus of interest of the different stakeholders, the issues of ethics and privacy for learning analytics emerged as a major theme. From the autumn of 2014 LACE together with SURF SIG Learning Analytics, and other EU projects like Open Discovery Space, WhatchMe, LEA’s Box, and PELARS, have organised workshops and discussions on EP4PA as separate events and as part of conferences in Zagreb, Amsterdam, Bolton, Paris, Graz, Osaka, and Helsinki.

During these events participants with backgrounds in teaching and learning, research, and ICT were asked to contribute their issues through filling out a Google form and in other ways. The following list of questions is an excerpt of the issues raised grouped in five categories, Privacy, Ethical Issues, Data Issues, and Transparency Issues.

A) Privacy

General
- Where is the boundary on data use for learning analytics (courses, grades, LMS, GoogleDrive, library system, residence halls, dining halls, …)?
- Where is the balance between benefits offered by LA/EDM and the protection of privacy?
- How can anonymity be ensured (especially in small groups) and how can it be systemised?
- How efficient can privacy protection be and how can the efficiency be measured?
- What are the dangers of learning analytics?
- How complete and permanent a picture do our data provide about students?
- What responsibility comes with ‘knowing’?
- Are existing legal frameworks, e.g. Data Protection Act, sufficient in protecting students’ privacy?

Access
- Who has access to data about students’ activities?
- What data should students be able to view, i.e. what and how much information should be provided to the student?
- To what extent do we provide students the option to update their data and provide extra (possibly qualitative) data?
- Do students have the right to request that their digital dossiers be deleted on graduation?
- Who has access to combined (e.g. from different courses) student data in what format?

Institutions / management
- What impact do privacy concerns have on an institution’s management and how can they be dealt with?
- What are the implications of institutions collecting data from non-institutional sources (e.g. Twitter)?
- Does the administration let students know their academic behaviors are being tracked?
- How much information does the institution give the teachers?

Consent
- When do students have to give consent to their data being collected and analysed? When don’t they?
• Should students be allowed to opt-out of having their personal digital footprints harvested and analysed?
• If we outsource the collection (and analysis) of student digital data to companies, do students need to give consent?
• Do open courses need consent?

Open data sources
• Is there a difference between data collection and analysis in open vs. closed courses?
• Do open courses need terms of use?
• Are social media data public and are they open to analysis?

B) Ethical Issues
• Where is the boundary on data use for learning analytics (courses, grades, LMS, GoogleDrive, library system, sports, residence halls, dining halls, …)?
• When is it ethical to show students the likelihood of their success?
• Should cloud services be used even if an institution does not permit such usage due to privacy or licensing concerns?
• Are there any circumstances when collecting data about students is unacceptable/undesirable?
• What amount of resources should the institution invest in students who are unlikely to succeed in a course?
• What obligation does the student have to seek assistance?
• What responsibility comes with ‘knowing’?
• Do students have a right to see results from external analyses?
• Should data from a platform used for corporate training of employees be made available to the employer?
• Is there a difference in collecting and analysing data from open or closed courses? Is consent needed?
• Is it essential to tell students about their activities being tracked and analysed?
• How can students be convinced to accept that their activities are being tracked and analysed?
• Are social media data public and are they open to analysis?
• Does the quantification of performance always have a positive impact?

C) Data Issues
• Are interconnected datasets a threat to personal and democratic principles?
• How much data sources should we combine to use in learning analytics? (all available, only relevant, only for one task?)
• Who is in charge of, i.e. who owns, the collected data?
• Who is accountable for personal data management?
• Who has access to what data and when?
• How complete and permanent a picture do our data provide about students?
• Are bigger data sets always better and do they provide a more complete picture than small data sets?
• What are the concerns when outsourcing the collection and analysis of data? Who owns the data?
• Is there a difference when dealing with data from open or from closed courses?
• Can students be given the control of their own data (privacy by design)?
• Is opting-out an all-or-nothing approach?
• Who has the responsibility over digital technologies, e.g. in case of erroneous feedback?
• How can reuse of collected data for non-educational needs (e.g. finance, insurance, research) be prevented? Or is it no problem?

D) Transparency Issues
• What should students be told about the data collected on their learning?
• What data should students be able to view about themselves?
• Does the administration let students know their academic behaviours are being tracked?
• What and how much information should be provided to the student?
• How much information does the institution give the teachers?
• To what extent should students have access to the content of their digital dossiers, who have access to these dossiers, and what it is used for?
• How can students be convinced to accept that their traces are tracked by tutors and administrators and is it really essential to tell them?
• What should be the communication and legal strategy to our students? (esp. considering that physical privacy via location beacons could be more than an issue)
• Do users and their parents know what data is collected and why?
• Are institutions open about secondary use of the collected data?

Ethics and privacy discussions are an ongoing activity within the LACE community, and it is too early to conclude with a clear plan of actions. In addition to giving input to development of policies for ethical use of student data for learning analytics there is a need to see what pedagogical, organisational and technical consequences the ethics and privacy concerns will have for bringing learning analytics out of the research laboratories into the daily practices of learning, education and training.

Observations from the workshop questions
Under the macroscopic issue of privacy, we distill the following as being relevant to this report
• The need to understand appropriate boundaries for data combination are identified as being significant; systems beyond the obviously learning-specific are called out: social media, dining, ... . This point is a recognition of the need for contextual integrity.
• The trade-off between benefits and privacy protection is identified. This idea could be developed further to be a conceptualisation of the problem as a balance between benefits and a risk to privacy, to account for the uncertainties in privacy protection.
• There is concern about the effectiveness of, and limitations to, privacy protection methods.
• Transparency, accountability, and control are seen as important facets of approaches to deal with the issue of privacy protection. The binary opt in/out model is challenged.
PART B 2 European Project Case Studies

This section gives an accounts of aims, requirements, and issues from a selection of European projects, both nationally-funded and EC-funded, and private enterprise.

2.1 Designing a national learning analytics system – the case of Estonia

Estonia is with a population of 1.3 million one of the least-populous states in Europe. With support from the Estonian Ministry of Education the Centre for Educational Technology at Tallinn University has designed learning analytics solutions for two nation-wide systems: a community-based digital learning environment eDidaktikum for teacher education and an Educational Cloud aggregating digital learning resources in secondary education. Dr. Kairit Tammets reports.

In developing learning analytics capabilities for the two national systems, the Estonian stakeholders wish to use learning analytics in a evidence-based policy making process, for example learning what kind of digital learning resources purchased with government funds are actually used by the schools. The aim is to develop learning analytics solutions that are pedagogy-driven and derived from research questions. Estonian stakeholders perceive that the value of learning analytics applications would be higher if they would be part of policy-making at the level of Ministry of Education and Science and used for educational research. The aim of the design has not been to create dashboards and giving feedback to learners and facilitators, but to make sense of data stored in national systems in order to understand how learning happens in digital learning ecosystems and to support Estonian educational research in general.

The following stakeholders are currently involved: Students (teacher education students at the university and secondary school students); Teachers/lecturers (teacher education lecturers and secondary school teachers); School leaders, administrators (school leaders from secondary schools, curricula managers and teaching assistants in teacher education context at the university); Publishers (textbook publishers, publishers of digital learning resources, owners of the repositories of digital learning resources); Policy makers (policy makers from the Ministry of Education and Science, responsible teacher education and policies regarding digital learning resources); and Researchers.

Two national systems have been targeted for learning analytics:

eDidaktikum

Estonian education is driven and derived from policy makers. The development of the system has followed participatory-design principles with the aim to bring the end-users closer to the development process. eDidaktikum provides a personal space for creating competency-based portfolios, accessing personal blogs and managing files. There is also a community space where learning materials can be created and shared, discussions held and task management organised. Learning analytics is implemented for supporting students and facilitators by giving feedback about course activities and progress. Also learning analytics data will be used for investigating online behaviour of students, and for evidence-based decision-making.

Educational Cloud

The Estonian Educational Cloud is a set of interoperable services that aggregate and publish metadata of digital learning resources stored in different Estonian educational repositories. Through the Educational Cloud, students, teachers, school leaders, and parents have access to variety of
digital content, stored in the repositories of Estonian educational organisations and publishers, and other web-based services such as YouTube, SlideShare, LearningApps, etc. Users can create web-based collections of those resources, and add comments and other user-generated content. As the resources are accessed via the Cloud, even if they are stored outside of the Cloud, monitoring the interactions between learners and content becomes possible for learning analytics purposes. Data will be used for giving feedback to users about usage of digital learning resources. Also school leaders can use data for making decisions about investments in (commercial) digital learning resources and their cost.

2.1.1 Requirements
The Estonian learning analytics solution is based on the following requirements:

Access to data
Data collected from one system (e.g. eDidaktikum) will be integrated with data collected from user surveys and other systems (different information systems, other LMSs). Quantitative data collected from clicks on content and time spent on activities will be expanded with qualitative content and semantic relations in the system. In principle, demographic data, activity data, and data from surveys can be shared with other interested parties provided the data are being de-identified.

Privacy and Consent
Privacy and ethics is a little explored area in Estonia. The implementation of learning analytics is still quite modest (mainly EU funded project initiatives and a few state level initiatives). Currently, the organisation that hosts the system acts as the data owner, also being responsible for data collection through tracking solutions. It is not clear who in Estonia should be responsible for appropriate and effective use of learning analytics, i.e., who could decide what data is collected and used for learning analytics. One could not say that management of schools, universities, and ministries is actively taking part in decisions concerning ethics and privacy of learning analytics. These decisions are done within research and development initiatives at the universities, and those teams are at the moment the de facto decision makers in control of analytics processes and regulating the ownership of data. At the management level of the institutions and at the ministerial level one could doubt that there is awareness of data about learners being collected or stored.

Currently the aim is to inform the students about what kind of data is being tracked and analysed, for what purpose, and by whom. Furthermore, students should be informed about what kind of analysis will be done, and with whom the visualisations will be shared. Currently consent has not been asked for collecting and analysing the data, even if it is mainly used in anonymised form and not used by third parties.

In the future, Estonia aims to give students some possibility to opt out of having their data collected, harvested, analysed and visualised. This might reduce the accuracy of learning analytics in the specific course context. Therefore, students will be educated beforehand about the consequences of withdrawal and there will be attempts to diminish the threats related with privacy concerns. The solution is not to leave the data in a closed repository under the control of one organisation. Discussions around trust, privacy and ethics should be initiated in Estonian learning analytics community soon. The discussion around open and linked data should be initiated as well – who thinks he owns linked data and who actually owns it?
Data access by third party
For the moment, very few people are accessing the collected data. Also it could be said that the research institute being part of the public university is responsible for guaranteeing ethical usage of data and control of data. People who would like to get the access to the data have to have an agreement, which sets restrictions of using data for non-educational purposes. Agreement between data holder and third parties will be formulated, regulating the data usage by third parties. Additionally, researchers will sign an agreement that sets restrictions for using data only for research purposes. In any case, only de-identified data will be shared for reuse purposes. User IDs will be regenerated in order to diminish the chances to map IDs with real users. In user-generated content user names and course names will be stripped - context will remain, but username will be replaced, e.g., with “User1”, “User2”, etc.

2.1.2 Technical Solutions
In developing Estonian learning analytics solutions we focused on the xAPI specification. The aim of using xAPI was to acquire practical experience on how to exchange data between different educational systems. The Estonian educational system is moving towards a distributed learning ecosystem that consists of different tools and technologies, and the challenge is to analyse the data about the same learners in different formats. As for medical information at the moment, we assume that learner as well is interested in accessing data collected from different educational systems from one place.

Via xAPI statements the following information will be communicated:

- Creating the content - learning resources, blog posts, discussions, collections, assignments;
- Commenting and replying - blogs, discussions, learning resources, assignments;
- Annotating and bookmarking of content;
- Remixing of the content (e-textbooks) and reusing of the content;
- Mapping the content (of e-textbooks e.g.) with learning activities from curricula;

In the process of implementing xAPI several barriers were identified:

- Verbs: On the one hand we experienced that there are too many verbs that may mean similar things. There were situations when we felt that we could use this verb or that verb, both seems suitable. Same time we experienced that there are some learning activities without suitable verbs (in the context of e-textbooks e.g.). And finally, during the development phase, verbs changed - some were combined with others;
- Context: Here also we could not perceive a solid approach how to define the context in a statement - one could go this way or another. At some point it was experienced that developing statements were based on a “hunch”, which means it is hardly possible to guarantee interoperability between systems;
- Multilingual approach: This might not be a real barrier, however, we are still struggling with this issue. Statements are sent to the LRS in the same language as it is created. If one users uses Estonian and reads “materjal 1” and another users reads the same content, “material1” in English, two different statements for “different” resources will be created, although it statements are related with same resource.
• Lack of best practices: In practice in education, xAPI statements are a rather new phenomenon and it was not easy to find best practices that really work and would achieve what we planned in Estonia.

Learning Record Store
The xAPI specification details a LRS as a store of tracking statements communicated through the xAPI. The tracking tool exchanges the xAPI statement with the LRS, which is able to exchange data with the final application. In the Estonian case, Learning Locker is the LRS of choice, as it is the reference open-source LRS and offers possibilities for customisation. In the development plans of Learning Locker it has been drawn up ideas to develop a personal LRS space for learners to own their data, which is also relevant for us. In the first phase of our development, the idea is to provide possibilities of having personal spaces in institutional LRS’s. In this personal space user can set up what kind of data will be collected and analysed about the user, who can have an access with what purposes. User can also set up what kind of information that can be visualised about the user to which stakeholders. In Estonia it is considered important to inform the user of the consequences of adjusting data usage possibilities by third parties. The aim of giving the users control over their data is to enhance the sense of control over the learning analytics process, to reduce the concerns of misuse of data, and to raise the awareness of what could be done with data in different systems to support lifelong learning outside of educational institutions.

Although Learning Locker provides a dashboard to visualise data to participants and it was initially argument for Learning Locker, in the Estonian case, dashboards are developed separately. Learning Locker functions as intermediate between the tracking tool and the dashboards.

The dashboards are made available to the different stakeholders in different systems. Based on statements in Learning Locker LRS, visualisations are developed with Highcharts solution and sent back to platform.

Mainly open-source solutions are used for developing educational systems and solutions. The source code is available via GitHub repository and for free everyone to use. E.g., the eDidaktikum platform and its dashboard module are open-source, using open-source LRS with xAPI statements, and the code is freely available for interested parties.

Observations from this case study
• The main interest with the Estonian learning analytics system described in this case study lies with the needs of the government to make better use of educational data for evidence-based policy-making. This requires access to data from outside the learner’s home institution.
• The Estonian Educational Cloud is a national service, which serves as a data hub and could potentially be a integration point for future learning analytics services and for centralised access control. As of today the focus is on learning resource usage statistics for resource planning and management.
• In sharing data between IT systems the focus has been on APIs and the use of the xAPI specification for expressing learning activity records.
• The organisational interoperability aspects of data sharing are acknowledged; however, the specifics of ownership and control, and clarity on the appropriate norms for privacy policy is still an emerging field of debate in Estonia, both in government and in research.
• This reports also points towards further development of the national architecture for
learning analytics, mentioning the interest to explore personal Learning Record Stores as part of Estonian learning analytics system.

2.2 PELARS – a European project investigating practice-based learning

PELARS\(^{33}\) is EU funded FP7 STREP project focused on investigating multimodal learning analytics for practice-based learning that is engaged in secondary schools and university education. The project has ten partners that include universities, NGO’s, and SMEs. The project collects data from ambient sensors (computer vision, smart furniture, data from programming tools) and user-generated content from learners (text and rich media). Dr. Daniel Spikol reports.

The data input to PELARS consists primarily of logged data, which comprises both the ambient sensors and the learner activities. Learner logging captures the different objects on the smart furniture, learner actions in small group work, and learner interaction with the hardware and software for the physical computing components. In concrete terms, PELARS will initially gather snapshots of the student’s programming code, the number of active hands in the learning environment, and the identity of objects that are being manipulated. Additional types of data will be collected as the PELARS project moves forward. The learner-generated content is captured via users’ mobile devices, tablet computers, or laptops and consists of text and rich media.

The system uses this information to compute additional information that can include: objects usage sequence, “hot” objects which are heavily manipulated, working and coding periods, and correct relative positioning of the different objects according to their type. The project is ongoing and additional metrics will be defined based on the design of the system for learning analytics that includes logic, reasoning, and visualisation.

\(^{33}\) [http://pelars.eu/](http://pelars.eu/)
Parties

PELARS consortium is intended to serve researchers undertaking formal research studies. It will be possible for researchers to use the PELARS as a private repository, with total control of who else may access the data they deposit. The project may make data-sets openly accessible, or may allow access requests.

Motivation

The motivation for sharing is to investigate how multi-modal data from ambient sensors and user-generated content can be used to support practice-based learning scenarios with tangible objects with digital components for learning analytics. Since capturing the data requires specialised equipment, the idea is to maximise the researcher-access to this data to get sufficient return on the investment.

Non-Technical Platform

The PELARS approach is built around a European collaborative/cooperative model and governance is member-led. Each member institution is required to follow its normal institutional approval process for human subject research. PELARS is working down from EU policies on data handling and privacy in conjunction with the different national policies. The majority of the data sharing is between
partners in a research context for the project, but future exploitation plans and sharing of the data-sets require a scalable solution for data policies.

**Technical Platform**
The storage is currently based on relational database pending continued development can be extended and with different software techniques. The storage is inherently structured and relational for the high-level parts with unstructured storages for the temporal types of data.

**Observations from this case study**

- Data sharing within a research context is governed by well established research ethics and procedures. However, when research results are commercialised there is a need for a more robust data sharing policy. It might be recommendable to include these issues also in research projects as the resulting solutions may be impacted by a more realistic and more complicated usage scenario.
- In the case of the PELARS project, there are ingredients like ambient sensors that points toward future learning scenarios with interesting data sharing challenges related to privacy. The inferences that may be drawn from this kind of sensor data may be less easy to anticipate, making it harder to understand the risk from disclosure.
- According to this report PELARS rely on a bespoke relational database for storage. When interoperability with other systems gets more focus the project would have to explore current specifications for exchange of activity information, and may find that arbitrary design decisions make this difficult.

### 2.3 LEA’S BOX – a European project creating a Learning Analytics Toolbox

*The LEA’s BOX*\(^{34}\) project aims at making educational assessment and appraisal more goal-oriented, proactive, and beneficial for students, and at enabling formative support of teachers and other educational stakeholders on a solid basis of a wide range of information about learners. Dr. Michael Kickmeier-Rust reports.

LEA’S BOX is a learning analytics toolbox that is intended to enable educators to perform competence-centered, multi-source learning analytics tailored to the very concrete demands and requirements of teachers and learners and based on:

- the foundations of sound psycho-pedagogical models,
- intelligent model-based reasoning services, and
- innovative visualisation techniques.

Therefore, the project is going to build upon the significant and well acknowledged body of existing work in the field of learning analytics and aims to enrich that with two distinct advancements:

- Contribute reasoning algorithms and services on the basis of valid competence-centred psycho-pedagogical frameworks such as Competence-based Knowledge Space Theory (CbKST) and Formal Concept Analysis (FCA).
- Contribute novel approaches to visualising activity/performance/achievement data by utilising methods such as structural Hasse diagrams as well as advancing Open Learner Modelling techniques.

\(^{34}\) [http://www.leas-box.eu](http://www.leas-box.eu)
At the heart of such a project is the treatment of sensitive personal data. The LEA’s BOX project complies with the European Group on Ethics in Science and New Technologies\(^{35}\), the ethics code of the Council of European Social Science Data Archives\(^{36}\) and with data protection legislation in the member states where the research is carried out.

Specifically, for the LEA’s BOX system, the following data protection and ethic aspects are important:

- Informed consent needs to be given using an opt-in policy before data collection takes place (whether this is electronic or otherwise).
- Users have the right to withdraw their permission for the data to be used at any time, therefore the mechanisms handling data collected electronically about students’ and teachers’ use of LEA’s BOX must have facilities/processes to exclude specific users’ data within analysis.
- Even if consent is not given for data to be collected, users should still have access to the analytic facilities of LEA’s BOX.
- For parts of the system where cookies are used, an explicit agreement from the user is required for proceeding (e.g. using a modal dialogue), if the cookie will persist after the end of a session of use.
- All consent requests must be in the local language and contain plain wording.

In addition, special data regulations for the project\(^{37}\) as a research activity cover:

- Results of LA/EDM research must not be reported back to teachers and students in real educational settings unless the analytics and data mining algorithms are satisfyingly valid, reliable, meaningful, and relevant. A satisfying quality is subject to scientifically sound validation studies, of which the results must be part of an informed consent.
- Data sent to or obtained from components of the system must be fully anonymised in cases where no clear and reasonable need is given to pass personal user data to the system. This is for example the case when standalone learning software is using the project’s system. The external application should not pass any personal data but only a token to LEA’s BOX.

**Observations from this case study**

- LEA’s BOX has identified a number of design requirements, e.g., opt-in policy before data collection, right to withdraw, consent request dialogue in local language, etc.
- Social science research ethics is guiding the approach to personal/sensitive data in addition to requirements arising from data protection legislation.
- The project has also published a separate deliverable on Privacy and Data Protection Policy, motivated by the wish “to avoid a situation where we know we should address the topic, however, push it aside because of not knowing how”. It seems that there is still work to be done to turn soft ethical requirements into technical specifications of how the users’ privacy and data protection should be maintained in the specific tools.

\(^{35}\) [http://ec.europa.eu/epsc/ege_en.htm](http://ec.europa.eu/epsc/ege_en.htm)

\(^{36}\) [http://www.cessda.org](http://www.cessda.org)

\(^{37}\) The full ethics and data treatment policy of the LEAS-BOX project is described in the public deliverable D2.3: Privacy and Data Protection Policy, available at [http://css-kmi.tugraz.at/mkrwww/leas-box/downloads/D2.3.pdf](http://css-kmi.tugraz.at/mkrwww/leas-box/downloads/D2.3.pdf)
2.4 WATCHME project – case

The WATCHME project aims to improve workplace-based feedback and assessment and professional development by means of learning analytics. Suzanne Schut and Joyce Moonen report.

In WATCHME, data that is stored in the electronic portfolio system (EPASS) by a learner is owned by them. The learner can share the portfolio data with assessors. Sharing can be established at the level of a single assessment or of the complete portfolio. Educational institutions can be granted access to the portfolio data of their learners, i.e. via EPASS the portfolio of every student can be viewed individually. In EPASS, this is only allowed for undergraduate students. In all other situations, information is only obtained on more general levels via a Management Information Module in the system. The exact content of the information can differ and is custom made to fit the requirements of the institution. As an example, this information can contain the number of forms validated in one hospital, an overview of learners that completed their education and the corresponding date, etc. Any information that is in the portfolio can be queried, but the system administrators define the available queries per institution.

Within the WATCHME system, data is shared between the electronic portfolio and the student model (SM). The sharing is established using APIs. The data that is sent from EPASS to SM consists of student scores received through workplace-based assessments that are based on an underlying competency framework, and structured in EPAs (Entrusted Patient Activities). Results from the student model (user interface including results from the learner analytics model) are sent via an API to EPASS, stored in a JSON array.

All users have a unique identifier. As long as a student (user) does not disable the use of the student model in EPASS, the same identifier is maintained. When a student opts out, the identifier is deleted from EPASS. From that moment, there’s no reference remaining to that student and thus privacy is maintained.

Use for Research

If EPASS data is used for research purposes, a dataset is created only containing the information that is necessary in the research. This set is created by a system administrator of EPASS, i.e. a person that has access to the database to take care of the administration and functionality of the system. There are no other persons able to query the database directly. The extracted dataset is coded so that no information can be traced to a particular user. This dataset is sent to the researcher via DataVerse, which is also used for backup and storage. The system administrator saves queries and the extracted databases on a secure server and/or an encrypted hard drive.

Observations from this case study

- The system implemented by the WATCHME project may be used for research. The developed solutions, however, are intended for workplace-based feedback and assessment and will meet the data sharing challenges as soon as the tools are brought out of the laboratories.

38 http://www.project-watchme.eu/
2.5 Building a cross-sector service platform – the case of Connect, Norway

In Norway, the higher educational institutions as well as primary and secondary schools have chosen a federated identity management system (Feide) based on the concept that services rely on user authentication at the user’s home organisation obtaining information about the user for its authorisation decisions. Tore Hoel reports.

Feide uses this federated approach to guarantee that each party remains in control, sharing only the necessary information to be authorised to access services. The home organisations, e.g., the school or university, register and authenticate their members; the service providers define their access rules, which are based on attributes provided by the home organisation. These attributes may be name and role in the organisation; and what information that is communicated is approved by the user. To meet new usage scenarios, however, the service providers want access to a wider range of attributes, information about the user before she logs on, and a Feide service that is easier to implement for mobile applications. This has led to the next step, Connect, a new service platform that could be used for learning analytics providers. The system is developed and in 2015 offered as a pilot by Uninett, the company that operates the Norwegian national research and education network.

Connect is a service platform brokering between services and systems that have information about users and a range of specialised educational services, e.g., publishers offering digital learning resources based on curricula and individual progression. The service providers will connect to and retrieve information through standardised APIs. The authentication of users will go through Feide and Guest IDs, but also potentially through the national ID portal operated by the Agency for Public Management and eGovernment.
Figure 4: Data owners offer information through standardised interfaces (APIs) in Connect.

Figure 5: Detailed architecture of Connect
Ironically, learning analytics is not part of the use cases informing the design of Connect. This might be explained by the fact that it is still early days for learning analytics in Norway. It could also be that the vendors would like a more centralised system or think the adoption of Feide is too slow within the school sector (see the case of the Norwegian learning analytics vendor, section B 2.7 A Norwegian learning analytics vendor’s perspective on data sharing and interoperability).

From a learning analytics architecture point of view, what makes Connect interesting is the way the data protection challenges are solved. First, a middleware layer under the control of educational authorities is put between the service providers and the users. In order to pass the API gatekeeper a 3rd party provider must sign a data processing agreement. There are four actors concerned about data protection, the end users (employees, students, etc.), host institutions (universities, school owners and other institutions), Connect, and 3rd party service providers. The host institutions, Connect and the 3rd party organisations sign data processing agreements containing instructions to Connect from the host organisations, e.g., about the usage of end user information, how end user rights will be maintained (access to information, rectification and erasure), and how the information will be secured against unauthorised access, exposure, change, damage or loss.

These actors take part in two usage scenarios: Usage initiated by the individual student, and “mandatory” use, where the host institutions decide which external services should be used. In the former scenario, informed, voluntary and explicit user consent is required for data access by the host organisation, Connect, and the 3rd party service provider. This consent may be withdrawn at any time, leading to deletion of information on the particular user by Connect and the 3rd party service (however, not by the host organisation). In the latter scenario, “mandatory use”, the host organisation decides that all or groups of end users should use 3rd party services or services they offer themselves. The legal requirements lie with the host organisation. If individual consent is not possible, other requirements must apply, i.e., the balancing of interests expressed in European and Norwegian law, e.g., purpose limitation.

Observations from this case study

- The system described in this case study aims at creating the middleware layer asked for in other case studies in this report, e.g., the Norwegian vendor and the Norwegian publisher.
- With a national system for authentication and authorisation of users there is a chance that vendors will see their interest in sharing data, providing there is a mechanism to manage the complex transactions. Connect aims at delivering such a mechanism through openly defined APIs.
- The shared approach to data processing agreements reduces cost of doing this separately, introduces efficiencies into the operations of the 3rd party providers and should reduce risk.

2.6 Requiring an open architecture for learning analytics – the case of JISC plans for the UK

Jisc is a registered charity which champions the use of digital technologies in UK education and research. Working in partnership with its customers, Jisc uses a “co-design” process to ensure that funding is allocated to the most pressing priorities for universities and colleges. Niall Sclater reports.

39 http://www.jisc.ac.uk
2.6.1 Introduction
In 2014 learning analytics was identified as a key area for development, and representatives from higher and further education identified and ranked the priorities:

Priority 1: Basic learning analytics solution. A “freemium” solution for further and higher education institutions would allow them to gain experience and eventually progress to a more advanced toolset. It would be based on existing products from institutions, vendors or open initiatives. The solution would require the implementation and/or development of open standards for analytics and APIs enabling multiple tools and data sources to connect.

Priority 2: Code of practice for learning analytics. The potential benefits of learning analytics are well recognised but there are also possibilities for misuse. A code of practice was considered to be the best way for institutions to assess and deal with the many legal and ethical issues arising.

Priority 3: Learning analytics support and networks. The group also prioritised the development of a support and synthesis service around the use of learning analytics to share expertise in:

- Technical methods – the nuts and bolts of learning analytics such as what systems and data institutions are using
- A learning analytics cookbook – with recipes for the use of data and metrics – documenting successful implementations at universities and colleges
- Synthesis and analysis – giving a high level overview and showing trends across the sector
- Networks – building networks of institutions keen to share experience both at a basic and advanced level

2.6.2 Outline architecture
Responding to Priority 1, Jisc is procuring the elements of the basic learning analytics system\(^{40}\) and has developed an architecture consisting of a number of components, which institutions will be able to select as required. The key components (see figure 1, below) are:

A **learning analytics processor** – a tool to provide predictions on student success and other analytics on learning data to feed into student intervention systems.

A **staff dashboard** – a presentation layer to be used by staff in institutions to view learning analytics on students. Initially this presentation layer will be focussed on the learner but dashboards for managers, librarians and IT managers could also be developed.

An **alert and intervention system** – a tool to provide alerts to staff and students and to allow them to manage intervention activity. The system will also be able to provide data such as intervention methods and whether these were successful or not to be fed into an exemplar “cookbook” on learning analytics.

A **student app** – based on requirements gathering\(^{41}\) with staff and students. Integration with existing institutional apps for students will be supported.

\(^{40}\) [http://analytics.jiscinvolve.org/wp/2015/04/04/explaining-jiscs-open-learning-analytics-architecture](http://analytics.jiscinvolve.org/wp/2015/04/04/explaining-jiscs-open-learning-analytics-architecture)

\(^{41}\) [http://analytics.jiscinvolve.org/wp/2015/03/18/gathering-requirements-for-a-student-app-for-learning-analytics](http://analytics.jiscinvolve.org/wp/2015/03/18/gathering-requirements-for-a-student-app-for-learning-analytics)
A learning records warehouse – a data warehouse to hold learning records gathered from a range of institutional data sources. We will define an output interface and also support integration with a common set of institutional systems.

A student consent service – allowing students to manage permissions relating to the use of their personal data.

Figure 6: Jisc's basic learning analytics architecture

The basic learning analytics system will be offered as a hosted, multi-tenanted service, keeping institutions’ data separate from that of other institutions. Some have already expressed a preference for hosting the systems themselves – this will also be facilitated. The systems which feed the learning records warehouse will also, in practice, be realised as a combination of locally-hosted and Software as a Service. Furthermore, the procurement process is structured to allow different suppliers to provide their own learning analytics components. Hence, the realisation of the architecture implies formal agreements about transfer of data about a single user between systems operated by multiple entities.

2.6.3 Why an open architecture?
While the various elements of the basic learning analytics solution can be proprietary or open source, the architecture itself is being developed using open standards. An open architecture aims to maximise opportunities for uptake and sustainability of the basic learning analytics solution. The following benefits are anticipated:
1. Having multiple products fitting the architecture helps it to achieve a critical mass.
2. Institutions can select only the components they require without having to invest in a single large monolithic system.
3. Tools can be replaced relatively easily if a better or cheaper component evolves.
4. An open architecture which is well understood and adopted should allow the market to develop, with new tools able to be integrated to enhance the possibilities of learning analytics.
5. A mixed economy of open source and vendor driven products facilitates healthy competition, drives innovation, and allows real choices to be made by institutions; it also stops a single vendor from developing a monopoly.

Jisc believes that the long-term benefits of an open architecture outweigh the cost difference relative to a turn-key solution and the effort required to realise a multi-party distributed system. In addition to addressing technical interoperability using open standards, the implementation of an open architecture will also require attention to legal, policy, and process aspects of data exchange, sometimes referred to as organisational interoperability. Action against priority 3 (Learning analytics support and networks) will be an important tool to develop the necessary organisational interoperability.

2.6.4 Ethical, consent and control issues
Whenever learning analytics is discussed in detail with staff and students ethical and legal issues are inevitably raised. Jisc’s co-design process identified the need for a code of practice to help institutions implement measures to address the fears of students and staff and to accelerate progress towards the real possibilities of using learning analytics to aid retention and student success. Some of Jisc’s stakeholders considered it an essential step at their universities and colleges before they could make progress in the area. In addition to saving the cost of each institution developing their own code of practice, collective alignment around a common approach has also addressed some of the requirements of organisational interoperability.

The first task was a review of the literature\(^42\) in the area. Material was drawn from 86 publications, including 16 codes of practice or lists of ethical principles from related fields. From this a taxonomy of ethical, legal and logistical issues\(^43\) was developed and refined in collaboration with Apereo and the LACE Project. This categorises the issues under the headings: Ownership and control, Consent, Transparency, Privacy, Validity, Access, Action, Adverse impact and Stewardship. Each issue was given a name, a question which embodied it, a type (logical, legal or ethical), a ranking of its relative importance, and the type of person or group which needs to take responsibility for it at an institution e.g. senior management, data scientist or student.

A Code of Practice\(^44\) was then drafted from the taxonomy under the guidance of an advisory group comprising representatives from universities, colleges and the UK’s National Union of Students. This brief document was released for consultation to a wide range of institutions, groups and

\(^{43}\) http://analytics.jiscinvolve.org/wp/2015/03/03/a-taxonomy-of-ethical-legal-and-logistical-issues-of-learning-analytics-v1-0/
\(^{44}\) http://analytics.jiscinvolve.org/wp/2015/04/21/code-of-practice-for-learning-analytics-public-consultation/
organisations, and their extensive feedback incorporated. Jisc will be developing online guidance to provide additional support for institutions which wish to implement the Code.

**Observations from this case study**

- Although Jisc is developing a learning analytics solution, the IT systems will be multi-tenant SaaS. This means that, although the IT systems may be operated by third parties, the data from different institutions is operationally separated and Jisc has no privileged access. Hence, “solution” refers substantially to a common: specification, framework for licensing, and architecture.
- The adoption of an open architecture and with distributed systems and choice of components resting with the individual universities and colleges emphasises the importance of technical and semantic interoperability. The working assumption is that there will be sufficient similarities between the universities and colleges for organisational interoperability not to be an issue.
- A code of practice for learning analytics has been identified as a key enabler. The driver for this has been a need for institutions to understand how to navigate the concerns of staff and students, and it naturally leads to an approach to dealing with some data sharing concerns, as well as a range of other aspects of the conduct of learning analytics.

### 2.7 A Norwegian learning analytics vendor’s perspective on data sharing and interoperability

*How would a vendor serving the majority of the local authorities in Norway look upon data sharing and interoperability related to learning analytics? This small case study is based on an interview with the head of research and development of Conexus, Yngve Lindvig, public presentations and conversations with a number of sources within Norwegian education. Tore Hoel reports.*

In an open edtech industry meeting, Conexus, a leading Norwegian learning analytics company also targeting an international market, summarised the current limitations as follows: not all information is in the same system (the ecosystem is incomplete). The parts of the ecosystem Conexus has in mind are school administration systems, learning management systems, user-generated content, content from publishers, self-assessment, learning standards (curriculum goals), etc. The problem perceived by this Norwegian vendor is access to the user catalogue, i.e., identified students and their affiliation. There is no willingness by the other suppliers in the market to share access to their users, and the government has no unified, national user catalogue. (In Norway there is a federated identity management scheme called FEIDE⁴⁵ - however, vendors see this as a single sign-on system not yet fully implemented in all schools, with a number of shortcomings in terms of information attributes and platforms support.) Conexus wants a national system run by the government with a standardised API for exchange of data, allowing different vendors to connect, read and share data, making it possible to build on and enrich information that other vendors have contributed. This, however, implies that the government must take an active stance to create the basic infrastructure on which the companies can innovate. It also implies, it has to be said, agreements between vendors regulating how to share data. As of now, the different players on the market will not share data, without a guarantee that they will become the dominant player or gain new markets. According to the Norwegian company Conexus, there is no incentive in the market today for developing APIs and other services that make data sharing possible, and no support from the government. As a result,

⁴⁵ [http://www.feide.no/introducing-feide](http://www.feide.no/introducing-feide)
there is a danger that an international actor, like Facebook or Google, will be seen as the dominant sign-on service by the users, and thus be the hub and thus getting the upper hand in developing a whole range of educational services. The Norwegian vendor does not think this will be in a national interest of building support for this country’s curricula and local services.

How would Conexus’ ideal system look like? The company has developed a blueprint for a system they want to establish with or without the help from the government. The proposal is not public. However, it involves an “anti-portal” that gives relevant services depending on whom the student is and what services and vendors she is subscribing (or entitled) to.

Conexus sees the current limitations in today’s market summarised as follows:

- Not enough standards in the ed-tech industry
- The ed-tech industry is not cooperating
- The target is not the students learning but the government’s need for control, or researchers’ joy in finding correlations
- Researchers focus on algorithms or teaching, not the processes in an ecosystem
- The government focus on content and not structures
- Professional capital and capacity building is not the target for the tools and reports
- Too much data driven control or personalised tasks, not enough evidence based processes and learning to learn
- Not enough focus on the markers of future performance and formative assessment (Improve school performance by assessment for learning)
- Not enough trust in the relation between government, school leadership, teachers, students and the community

To move forward the Norwegian vendor Conexus communicated the following needs, asking for:

- International integration and production standards
- Governments that implement national user catalogues
- Governments that support initiatives that participate in ecosystems
- Governments that support incentives to cooperate
- Governments that produces a starting point by implementing national data open to a common eco-system
- Governments that focuses more on professional development than data driven control

**Observations from this case study**

- This case raises the need for coordination and standardisation activities. It point to the role of the educational authorities to be more active in levelling the field of competition between vendors of learning analytics systems.
- The importance of control of the user catalogue is underlined in this case: who controls the authorisation process controls the market.
- When companies move from a single-vendor solution and vertically integrated market to a broader market procedures for data sharing must be put in place. Engagement for educational authorities is necessary to make move forward and achieve national solutions.
- The industry may have a too simplistic view on how simple it will be to establish technical solutions for data sharing across vendors, tools and educational actors maintaining all aspects of data protection, privacy and trust.
• The industry must be willing to share ideas about architectures and solutions in order to make progress in building the middleware infrastructure necessary to achieve data sharing for learning analytics.

2.8 A Norwegian publisher with adaptive learning product cooperating with a US company

The Norwegian publisher Gyldendal Undervisning has launched its Multi Smart Øving platform\(^\text{46}\), using adaptive learning and learning analytics in cooperation with Knewton, a US adaptive learning company that has developed a platform to personalise educational content. Presented here are the data sharing challenges the Norwegian publisher is facing. Project manager Espen Tokerud reports.

When sharing information one has to be aware of the fact that it is the school owner, in Norway the municipality or county, that is responsible for what kind of student information is being stored and how it is used. As a publisher, we sign data processor agreements with the school owners. Ideally, these agreements should describe in detail what data we store, and how we use the data, so that the municipality could make a good judgement on behalf of the students. This is the school owners’ responsibility, and they will be audited by the Data Protection Authority. However, if controversy should occur because the agreement is not clear, we as a publisher could lose credibility in the market. Coming up with a data processor agreement is not a trivial task, and if we share information with other actors the picture will be even more complicated. This in itself could make us more reluctant to share data than we would be under different circumstances. And, if we later on find that sharing data would be beneficial, we would probably need to renegotiate the agreement, which might delay the solution we are developing.

Our learning platform is designed for ‘volume training’ and a lot of information is stored about the students’ results. From the very beginning we have had a constructive dialogue with the Data Protection Authority. They want us to think data protection from the very beginning, applying the principles of Privacy by Design. Consequently, we have thought through what data we want to store, why we want to store it, and how to inform the end users. Therefore, we do not store information we cannot justify from a clear need on behalf of the teacher or the student.

Person Identity Information (name, school and class) are stored in Gyldendal servers in Norway, while the single interactions, i.e., what the students have read, or right or wrong answers, are sent to Knewton and stored in their system. We get analytics back on students achievements and produce reports to the teachers. Already in this simple data sharing setup there are a lot of factors. For example, Knewton’s servers are in the US, which make it necessary for them to have a separate "Safe Harbor" agreement with EU. We also have to make some decisions about the data stored by Knewton. If a student would use another product from Knewton provided by one of our competitors data gathered using our products could be used to give the student a better service. However, we have chosen not to go for this option, because we want to be very explicit in the data processor agreement with our customers. We want to be definite in our statements on what will and will not happen with the student data.

\(^{46}\) http://www.smartoving.no
In the Norwegian market there are companies specialising in learning analytics that we might want to work with in the future. For instance, our application could feed data into a system that combines data on learner achievements from tests and exams reported in different national systems, to potential great benefit for teachers and school owners. However, to design a good setup for cooperation is not easy. We need to have a go from The Data Protection Authority, and again we are hesitating as the dialogue with our users on how data are used becomes more complex.

Another complexity related to sharing of data is the idea to design a continuous learning trajectory in mathematics covering all K-12. In that case there will be a hand-over of data from one school owner (the municipality) to another (the county).

To summarise, as developers we do not see any technical restrictions preventing us from sharing data. It happens today between us and our partners. Neither is compliance with the legal restrictions an insurmountable problem when considering the relationship between Gyldendal and a single customer, since we can explore these issues with the Data Protection Authority. The difficulty in this situation is more about being thorough in documentation and communication with the bodies responsible for managing the schools. The data protection regulation is, however, imposing a structure of agreements between the school owners and each vendor. As a consequence, if there are more vendors working on a vertical solution the legal situation could become quite complicated.

Legal compliance may be complicated and time-consuming but it is a relatively well-defined problem. The tricky bit is, in our opinion, related to trust. Sharing of data may seem like a good idea for school owners or principals, and may be conducted according to the law and best practices in IT security. This, however, may be perceived as a threat of data protection by students and parents. The more data sharing that take place, the more stumbling blocks and unforeseen circumstances there might be that could hurt our credibility and our business. In the meantime, as a publisher knowing our market we will be a bit careful with sharing of data as we learn more about this new field of learning analytics.

**Observations from this case study**

- The contractual complexity related to data protection multiplies when vendors start sharing data.
- Data sharing across legal domains poses challenges related to where the data are stored.
- There is a need for tools developers to work in close contact with data protection authorities in order not to be slowed down in the development process and avoid being discouraged from taking their product to the next level.
- How much data could be stored by vendors for later analysis is depending on the data processing agreements with the customers. It seems that the data minimisation aim in the data protection regulations poses a challenge for development of tools for learning analytics.
- Are school authorities mature enough in their understanding of data handling and use for learning analytics to engage with suppliers to develop effective data processing agreements?
2.9 School owners’ concerns – what is allowed according to the Privacy Protection Act? The case of giving advice to Norwegian schools

The Norwegian Centre for ICT in Education, an agency under the Ministry of Education and Research, have been preparing a response to different initiatives related to learning analytics. The following is an excerpt from some informational material in preparation (translation and editing by LACE).

The school owner (i.e., the municipality) should consider the following issues before they could start using learning analytics in their schools:

- **Reference to law**: Information about persons can only be used for learning analytics if the school owner can justify the use according to law.

The Norwegian Personal Data Act (Personopplysningsloven) gives the following provisions the school owner must take into consideration:

- **Purpose of use**: Personal data can only be used for the intended purpose of learning analytics, i.e., increased learning outcome for the students. How will the school owner make sure that the data gathered are only used for learning and not for other purposes, whatever tempting, e.g., controlling students and teachers?

- **Data minimisation**: Only the personal data relevant to learning can be used for learning analytics. Where is the line between data that are relevant to learning and data that are not relevant, but could be interesting to register and analyse?

- **Data accuracy**: Personal data must be of a quality fit for learning analytics that actually contributes to learning. How will the school owner control the quality of data, especially if the learning analytics involves many external systems and service providers?

- **Storage and deletion of personal data**: The personal data must be deleted when they do not contribute to increased learning outcome any more (unless a record must be kept for archival reasons).

- **Rights**: Students, teachers, parents (guardians) have the right to information about how the school owners handle the personal data used for learning analytics. They can also ask to see, have corrected or deleted their own personal data. How should a school owner make sure that the data are made available, particularly if the learning analytics involves a number of external systems or services?

The draft text concludes: If these questions are not clarified and answered satisfactory it is likely that the applied learning analytics will be both illegal and that the school owner will not be able to maintain the most central principle of privacy: The data subject shall be given control with and consulted about the use of her own data.

**Observations from this case study**

- A restrictive interpretation of data protection regulations may block the scaling up of learning analytics in schools, e.g., “personal data must be delete when they not contribute to increased learning outcome”.

- In particular, there is a need to discuss the implications for learning analytics of the data protection regulation aims of purpose of use and data minimisation.

- The rights of the end users of learning analytics applications are maintained by school owners and others who procure learning analytics systems. There is a need for guidance on
how data protection is handled in these system, e.g., when a number of external system and services are involved.

- Data protection authorities and school agencies need to work together to promote knowledge about data protection and privacy related to learning analytics.
Part B 3 Existing examples of data sharing for learning analytics

This section is based on “Learning Analytics Data Sharing – Current Examples 2014”47, i.e. a working document submitted for public comment that will be updated over time. Each example presented in the current section uses a repeated structure. The elements of this structure are as follows.

Aims and Context
What is the educationally-relevant purpose of the activity, in terms of the kind of knowledge that is gained from the data, or the kind of actions that will be taken as a consequence of the analytics? What breadth of learner contexts is included, including phase of education, geography, ...? The motivation for sharing, rather than the educational intent is considered separately, see below.

Data
What kinds of learner data are shared? This aspect is concerned both with the kind of activity or learner attribute is to be analysed, and also the nature of derived data and results of the analytical process.

Parties
What kinds of party participate in a sharing relationship, and what asymmetries exist?

Motivation
What motivates the sharing, what specific benefits is data sharing believed to deliver that would be difficult or impossible to achieve within a single educational establishment? What is the business or funding model that is founded on this motivation?

Non-technical Platform
What kind of policies, procedures, and contractual relationships exist?

Technical Platform
What kind of technical approach is used? This could include choice of core technology, cross-party architecture, data exchange formats, and the role for technical specifications and standards?

3.1 Pittsburgh Science of Learning (PSLC) DataShop

Aims and Context
The PSLC DataShop is a repository and collection of analysis and visualisation tools to support a research community in the field of intelligent tutoring systems (ITS), in particular those systems based on the idea of Knowledge Components, which have their heritage in cognitive models of psychology. Knowledge components are defined on the PSLC wiki48: “A knowledge component is a description of a mental structure or process that a learner uses, alone or in combination with other knowledge components, to accomplish steps in a task or a problem.” Its immediate aim is to support the continued development of ITS, but it should be noted that much of the data in DataShop comes from systems that used by many thousands of school students, in grades 6-12, in the United States of America.

48 http://www.learnlab.org/research/wiki/index.php/Knowledge_component
**Data**
The data input to DataShop consists of ITS log data, which comprises both the learner activity and the “tutor” (software) response. Learner logging captures responses to problems but can go down to the level of interaction with individual user interface components, and tutor logging includes presumed attainment of Knowledge Components. Data export fields are essentially the same as imported fields, but with the potential for roll-up (e.g. individual interactions rolled up to student-problem level) and associated computation of simple aggregate statistics. DataShop visualisation and reporting tools allow for general purpose manipulation and visualisation of attributes such as error rate and step durations at various levels of aggregation. They also allow for visualisation of learning curves, i.e. displays of the changes in student performance over time. These are based on the Knowledge Component paradigm.

**Parties**
DataShop is operated by researchers at Carnegie Mellon University (CMU) and is intended to serve researchers undertaking formal research studies. It is possible for researchers to use the DataShop as a private repository, with total control of who else may access the data they deposit. Depositors may make data-sets openly accessible, or may allow access requests. CMU appears to have no privileged position with respect to data access.

**Motivation**
The motivation for sharing is the more efficient development of knowledge by enabling multiple research enquiries to be based on a single data-collection activity. DataShop was established using grant funding from the US National Science Foundation. It is understood that, at present, it can be maintained with modest levels of funding contributed from research projects making use of the DataShop.

**Non-technical Platform**
The default presumption is that data-sets added to DataShop remain private to the depositor. Since DataShop is operated by, and for, a research community, primarily in the USA, the policies in place assume and require adherence to ethical research practice, including review by institutional ethics committees (IRB, Institutional Review Boards, in US terminology). DataShop reviews evidence provided by depositors that the data has been ethically collected and appropriate consents obtained, before permitting depositors to share it with select other researchers, or publicly. All data must be de-identified.

**Technical Platform**
DataShop is a centralised repository with import via XML and tab separated (CSV) files, and export of tab-delimited data. Importing data is a batch process, in which the uploaded XML/CSV files are submitted to a queue. The XML format, which is extensively documented as a public specification, the Tutor Messaging Format, provides for a richer data import. This specification has evolved over time, but remains specific to PSLC.

**Further Information**
The DataShop repository home page provides a listing of public data-sets:
https://pslcdatashop.web.cmu.edu/
The “help” page contains information on the technicalities, procedures, sharing policies, and reporting features of the DataShop:
http://pslcdatashop.web.cmu.edu/help
This information is summarised in a “cheat sheet”:
The Tutor Messaging Format is an XML specification for logging data:

3.2 Predictive Analytics Reporting Framework (PAR)

Aims and Context
The Predictive Analytics Reporting (PAR) Framework focuses on Higher Education student retention and completion, across a range of types of institution in the USA (two and four year courses of study, public and private, traditional and non-traditional institutions). It undertakes benchmarking, prediction, and work to understand the signs of risk vs. progress to completion. In addition to prediction, the aim is to support the identification of good practice in student retention through data analysis, shared models, and benchmarking across institutions.

Data
The PAR Framework uses a fairly small number of variables compared to all of the data that could be collected in a learning analytics scenario, although over 60 are collected. These come under headings such as demographics, course of study information, academic records, and institutional type/approach. Student-level predictions of risk, and aggregated benchmark data is returned to member institutions.

Parties
The PAR Framework is managed by the WICHE Cooperative for Educational Technologies (WCET), a non-profit organisation. They describe PAR Framework as a collaborative, which currently consists of 16 institutions. Each of these institutions contributes its data to a central database managed by WCET, and receives the results of student-level analysis on its own data; WCET maintains a team including data scientists. Benchmark data is available to all member institutions. WCET appears to not make its own use of the data, but to be a centre of expertise and provider of technology to the members.

Motivation
The PAR Framework motivations are two-fold: a) that there is a cost-saving in having a central analytics service with highly skilled staff, covering multiple aspects of expertise from data science to policy and HE practice; b) cross-institutional benchmark studies provide valuable information on effective strategies to promote achievement, engagement, and progress. Initial setup funding, and funding to support the initial 16 institutions was provided by the Bill and Melinda Gates Foundation. The PAR Framework is now a non-profit provider of learning analytics as a service, funded by through subscription fees by member institutional partners.

Non-technical Platform
The PAR approach is build around a collaborative/cooperative model and governance is member-led. Each member institution is required to follow its normal institutional approval process for human subject research (ethics committee/IRB) and the PAR team all have certification in human subject research. PAR indicates attention to data privacy, record security, and research integrity.
**Technical Platform**
The technical platform is essentially invisible; WCET clearly operates an enterprise database and appears to accept data in batch files. The PAR Framework publishes its set of common data definitions under a Creative Commons licence. These are not fully-specified, and are not mapped to data definitions from other sources.

**Further Information**
Information about PAR and their approach:
http://www.parframework.org
Data definitions:
https://community.datacookbook.com/public/institutions/par

### 3.3 Open Academic Analytics Initiative (OAAI)

**Aims and Context**
The Open Academic Analytics Initiative (OAAI) is a completed project, largely undertaken in 2012 involving two US community colleges and one Historically Black College and University in addition to the lead, Marist College, a private US higher education institution. It focussed on predictive modelling of students at risk of dropping out with a view to directing interventions, and researching different intervention strategies. The project is of particular interest because it considered the portability of predictive models, i.e. model sharing, both technical aspects and statistical performance.

**Data**
Just short of 30 data elements were used, of similar type to those used in the PAR Framework, largely relating to demographic details and academic performance, but with addition of grade-book and tool usage counts from Sakai.

**Parties**
This was a project with no ongoing data sharing.

**Motivation**
The particular motivation for data sharing was the investigation into the portability of predictive models across diverse academic contexts (but all US post-secondary).

**Non-technical Platform**
This was undertaken as human subject research; ethics committee (IRB) approval was gained in all institutions.

**Technical Platform**
OAAI is distinctive in that the project specifically set out to develop and deploy a system comprising a number of Open Source Software components. The principal point of interest for this case study is their use of an open standard, PMML (Predictive Modeling Markup Language), to export their predictive models.

**Further Information**
The OAAI project page outlines the project and its results, with abundant links to reports and technical documentation, as well as videos:
https://confluence.sakaiproject.org/pages/viewpage.action?pageId=75671025
A description of the data elements used:
The OAAI project has influenced the development of the Apereo Learning Analytics Initiative\(^\text{49}\), and a high level system architecture (below), which has in turn influenced the Jisc learning analytics architecture (section B 2.6 Requiring an open architecture for learning analytics – the case of JISC plans for the UK).

**Figure 7:** Apereo Learning Analytics Initiative High Level Architecture, the “LA Diamond”, by Aaron Zeckoski

### 3.4 Multimodal contextualised Learner Corpus Exchange (MULCE)

MULCE\(^\text{50}\) was a research project supported by the National Research Agency in France between 2007 and 2010, predating all of the other examples in this document. The MULCE team worked on requirements for research data to be shareable from a variety of perspectives and built a repository. It is now hosted by the LRL (Laboratory for Research on Language) in the MSH (Maison des Sciences de l'Homme) in Clermont-Ferrand and used for research on multi-modal interactions in language learning. “MULCE” will be used in the present tense to refer to the project and to continuing activity together. Since MULCE has operated for several years, albeit at a modest scale, it has learned some of the lessons of data sharing for a research community (see Further Details) including the practicalities of exchanging and re-using copora, structuring multi-modal data and contextual information, and ease of access for deposit or by analytical tools.

\(^\text{49}\) https://www.apereo.org/content/learning-analytics-initiative

\(^\text{50}\) http://mulce.org
Aims and Context
MULCE seeks to support an improvement in the quality of academic research in technology enhanced learning by broadening access to data-sets.

Data
There are two distinctive features of the MULCE repository: the variety of modalities of interaction, including virtual worlds and white boards as well as synchronous and asynchronous text, for example; a distinction between “global” datasets, which are the original form, and “distinguished” datasets which are the selected and transformed data used in particular research. MULCE is also concerned with capturing, in variously structured forms, metadata to describe the context from which the learner activity data was captured.

Parties
MULCE is, in principle open to any interested researcher to deposit data, although it remains necessary for MULCE team members to be involved in the process for technical and quality assurance reasons. The repository is open access for data download. The terms of use specify non-commercial use and note that the intended users are researchers and teachers.

Motivation
What motivates the sharing, what specific benefits is data sharing believed to deliver that would be difficult or impossible to achieve within a single educational establishment? What is the business or funding model that is founded on this motivation?

Non-technical Platform
The standard licence for use of data in the MULCE repository is Creative Commons Attribution Non-Commercial Share-Alike, although there are clauses which permit a depositor to add conditions. The terms of use indicate, but do not require (the word “should” is used), that users should alert the depositor if personal data is discovered. MULCE has clear expectations for the level of detail of the context of the data, including what might be called the learning design as well as the higher level aims of the course. Manual intervention from the MULCE team forms part of the deposit process.

Technical Platform
MULCE has taken a mixed approach to data exchange formats, preferring to use or be inspired by published interoperability standards while developing its own XML-based structures for expressing the activity and contextual information. They use, or refer to, for example: the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), IMS Content Packaging, IMS Learning Design, and IEEE Learning Object Metadata (LOM). OAI-PMH permits the harvesting of the metadata about resources (in this case datasets) so that they can be indexed by, and discovered through, other services. i.e. they embrace the idea of a distributed architecture for resource management and discovery. MULCE also developed software to support the de-identification of learners in text-based chat and forum exchanges. Source code is available, but it dates from 2006.

---

Further Information

3.5 Stanford Data Portal for Research (VPOL)

This is also sometimes known as the Stanford VPTL\(^52\) (Vice Provost for Teaching and Learning) Data Portal. The term “VPOL (Vice Provost Office for Online Learning) Learning Data” is also used, apparently interchangeably.

Aims and Context
The Data Portal is intended to be used for academic research without any presumption of research topic, within the bounds of what is possible with the data.

Data
Data from MOOCs operated by Stanford University is available, specifically from the use of three platforms: NovoEd, Coursera, and OpenEdX. The data describes assessment, forum, and video-watching activity. Interactions are available at an intermediate level of tracking (e.g. assessment responses and excerpts from video track-logs). A computed time-on-task is also available.

Parties
The data is available for academic research by scholars from Stanford and elsewhere. NovoEd data is only available to Stanford.

Motivation
The motivation is presumed to be a combination of Stanford University seeking to enhance its reputation and to manage exploitation of the data by its own academic staff (“faculty”). It appears to be entirely self-funded by Stanford, a well-endowed institution, so likely to persist in the medium term.

Non-technical Platform
Stanford’s MOOCs make it clear that data from learner activity will be used for research as part of its Terms of Service, which were revised following the 2014 Asilomar Convention (see Further Information). Access to the data requires assent to a data use agreement and formal assessment of all requests for data access by the VPTL Data Sharing Working Group, which reserves the right to revoke its consent. Requests are required to specify the project for which the data will be used and to limit use to that purpose. Disclosure to third parties is explicitly prohibited, as is publication of any personally identifiable information. Furthermore, data users are prohibited from attempting to re-identify individuals.

Technical Platform
The data is held in a MySQL database with a stanford.edu URL. Notes are available explaining access from Excel and SPSS. For OpenEdX and Coursera course data, the structure of the database is a replica of that used in the production platform. MOOCDb versions are also available for some

\(^52\) http://vpol.stanford.edu/research
courses; MOOCdb is a project founded at MIT to develop a platform-agnostic relational data model for MOOC data.

Further Information
Data use agreement:
https://stanford.box.com/dataset/agreement
Description of data structures:
http://datastage.stanford.edu/
Asilomar Convention, six principles agreed on by a meeting of scholars concerned with “the collection, storage, distribution and analysis of data derived from human engagement with learning resources”: http://asilomar-highered.info/

3.6 xMOOC platforms
xMOOC platforms are better described as shared delivery platforms, rather than shared data platforms, but they are included for three reasons:

- these platforms typically offer the usage data back to their clients;
- the scale and relative uniformity and simplicity of the delivery model\(^{53}\) makes the data attractive for learning analytics;
- MOOC data is generally very difficult to fully anonymise since forum posts often contain either direct identification of the sender, or sufficient traces to allow for identity to be inferred when combined with social media and other public-access data.

The initial round of xMOOC platform implementation focussed on the delivery platform and the provision of database dumps to their client/partner, while apparently making little central use of the data. Although little evidence of central use of the data is available, the current privacy policies are rather similar to Silicon Valley social media corporations, and markedly different from PSLC DataShop or the PAR Framework. Coursera usage data may, for example, according to their privacy policy, be shared with any business partner for research or to allow them to “share information about their products and services that may be of interest [to you].” Futurelearn includes statements with a similar sentiment and a similarly open-ended clause that “your information may be used by us and by technology partners and course and content providers chosen by us.”

The xMOOC providers’ initial focus on delivery platform rather than data use is now beginning to change, doubtless for a variety of reasons including: an ability to attend to less mission-critical ideas, now the core platforms are stable; a need to evolve sustainable business models; a need to compete on demonstrable effectiveness of the platform; an increasing interest from clients as they move from initial adoption into a more reflective stage of delivering MOOCs. This change is reflected in, for example, Coursera recruiting a Director of Analytics\(^{54}\) with a role including both platform-level analytics (e.g. A/B testing) and services to partner institutions (“university partners need the data insights we can provide them to advance their pedagogy”).

The European Multiple MOOC Aggregator (EMMA), a multi-lingual platform, includes learning analytics in its scope but is still developing its offering.

\(^{53}\) The vast majority of course use a combination of video lectures, short assessment quizzes, and online forums, while many also include mid-point and final graded assignments, sometimes deploying peer assessment.

\(^{54}\) https://www.coursera.org/about/careers/9d2e3d7a-e391-4197-bd5e-f2ed107bc800
3.7 Open data initiatives

This section does not describe a single example of a learning analytics data sharing as scoped-out in the introduction to this section, but considers two initiatives at the extreme end of data sharing: open data. LinkedUp is a project specifically targeting web data for education, which “aims to push forward the exploitation of the vast amounts of public, open data available on the Web, in particular by educational institutions and organisations”. The second initiative is the Open Knowledge Foundation (OKFN), which is also a member of the LinkedUp Project.

OKFN is concerned with furthering access to, and use of, knowledge in its widest sense, and through a number of means including campaigning, community support, consulting, and software development. It promotes publication of public sector spending data on the one hand, but also operates an Open Education Working group, about which it says “Open Education is much more than just OER and involves aspects like opening up relevant educational data and changing both institutional and wider culture”. In this case, and at present, however, “relevant educational data” is not thought of as the kind outlined in the introduction: data about people and their actions.

Putting to one side any speculation about what the Open Education Working Group may do in the future in the intersection between Open Education and Learning Analytics, the most significant contribution, from OKFN, of relevance to a discussion of learning analytics data sharing platforms is their open source data portal platform, CKAN, the Comprehensive Knowledge Archive Network. CKAN is used by numerous city and national government open data publication programmes.

CKAN has a number of features that may make it relevant to some realisations of learning analytics data sharing:

- Dataset metadata can be harvested, allowing for search across a federation of CKAN instances while data may be held back with access subject to separate agreement.
- Aside from being open source software, the CKAN architecture provides for “extensions” to be written without requiring source code changes in the main CKAN software. There is, for example, an extension to accommodate professional-level geospatial data in CKAN.

The LinkedUp project is responsible for the Linked Education Cloud catalogue of open data on the CKAN-powered datahub.io site. At present, this includes a small number of datasets of relevance to learning analytics, but outside the scope given in the introduction to this section, for example machine readable catalogues of learning objectives from the Achievement Standards Network. It includes only one dataset within scope, although aggregated and of tangential interest for learning analytics, which is circulation data from the University of Huddersfield library.
Further Information
The Linked-Up Project:
http://linkedup-project.eu/
OKFN mission and methods:
https://okfn.org/about/
CKAN open source software for data publishing:
http://ckan.org/
Linked Education Cloud:
http://datahub.io/group/linked-education

3.8 A counter-example: InBloom

Aims and Context
InBloom was a US non-profit with wide-ranging intentions to support progress tracking and delivery of individualised learning in schools. It wound up its operations in 2014, following widespread criticism and threats of legal action arising from the range of highly personal data that had been ingested into the InBloom databases.

Data
Since the InBloom ambition was a wide-ranging one, to support detailed progress tracking and individualised interventions by teachers, a very wide range of data fields were supported. Over 400 fields were supported, although it was a school district decision which fields were used, essentially the set of demographic, attendance, special needs, academic performance, and other data kept in school student records systems and reported to local authorities.

Parties
InBloom was essentially a technology provider to public education authorities, with each having a separate data store; to quote from their FAQ:

“States and districts are responsible for protecting student information. In accordance with both law and inBloom policies, inBloom is a third party service provider covered by FERPA, responsible for being an information steward to these public officials. InBloom provides each participating school district with its own protected storage space, so that district can continue to manage and control access to its own data, just as it always has.”

Motivation
The data sharing that InBloom enabled permitted the development of learning-related software by third parties, with the software having access to data that it would not normally have.

Non-technical Platform
The centre-piece of the InBloom approach was to provide the technical platform and for the school districts to be in control of the data. InBloom assumed responsibility for data security, including monitoring procedures, and designed the software to allow school districts to control data disclosure to third party providers of learning-related software.

Technical Platform
The InBloom platform was essentially a SaaS solution with support for bulk data ingestion of XML files as well as extensive REST APIs to allow third party developers to build new software applications. The data definitions were adopted from the Common Education Data Standards of the US Department of Education.
Further Information

The InBloom website appears to be no longer available, but snapshots from before the announcement of their intention to wind down may be found in the Internet Archive “Wayback Machine”:


There are numerous articles discussing the closure of InBloom, for example:


The Common Education Data Standards from the US Department of Education:

https://ceds.ed.gov/
Glossary of abbreviations and specialist terms

**API** – Application Programming Interface, the means by which software components exchange data or direct processing.

**CSV** – Comma Separated Values (also generalised to Character Separated Values), a simple textual representation of tabular data.

**ITS** – Intelligent Tutoring System.

**LMS** – Learning Management System.

**LRS** – Learning Record Store.

**Machine Learning** – the use of computer algorithms, as opposed to mathematical methods, to detect patterns in data such as cluster detection or prediction.

**MOOC** – Massive Open Online Course, commonly nuanced to xMOOC or cMOOC according to whether a conventional instructional style or a connectivist style of course organisation is used.

**REST** – REpresentational State Transfer, an architectural style for APIs that exploits the architecture of the web.

**SaaS** – Software as a Service, a kind of Cloud Computing.

**VLE** – Virtual Learning Environment, a regional term approximately equivalent to LMS.

**XML** – eXtensible Markup Language, a textual representation of structured information.

**xAPI** – Experience API, a specification for exchange of activity information.
References


Winne, P. H. (2006). Meeting challenges to researching learning from instruction by increasing the complexity of research. In P. H. Winne, K. Littleton, C. Constantinou, L. Mason, M. Roth, & R. Wegerif (Eds.), Handling complexity in learning environments: Research and theory (pp. 221–236).
### Version History

<table>
<thead>
<tr>
<th>Date</th>
<th>Notes</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-01-20</td>
<td>Collaborative online document created</td>
<td>Tore Hoel, HiOA</td>
</tr>
<tr>
<td>2015-05-01</td>
<td>Soliciting case studies, input text</td>
<td>Adam Cooper, Cetis, Tore Hoel, HiOA</td>
</tr>
<tr>
<td>2015-05-20</td>
<td>Internal review and comments</td>
<td>Dai Griffiths, Cetis</td>
</tr>
<tr>
<td>2015-05-01</td>
<td>Internal review and comments</td>
<td>Weiqin Chen, HiOA</td>
</tr>
<tr>
<td>2015-05-28</td>
<td>Amended version taking into account final review comments</td>
<td>Adam Cooper, Cetis, Tore Hoel, HiOA</td>
</tr>
<tr>
<td>2015-05-31</td>
<td>Cleared for submission to the EC Project Office</td>
<td>Maren Scheffel and Hendrik Drachsler, OUNL</td>
</tr>
</tbody>
</table>

---

### About this document

(c) 2015, Adam Cooper and Tore Hoel

Licensed for use under the terms of the Creative Commons Attribution v4.0 licence. Attribution should be “by Adam Cooper and Tore Hoel, for the LACE Project (http://www.laceproject.eu)”.

For more information, see the LACE Publication Policy http://www.laceproject.eu/publication-policy/.

### About LACE

The LACE project brings together existing key European players in the fields of learning analytics & educational data mining who are committed to building communities of practice and sharing emerging best practice in order to make progress towards four objectives.

- **Objective 1** – Promote knowledge creation and exchange
- **Objective 2** – Increase the evidence base
- **Objective 3** – Contribute to the definition of future directions
- **Objective 4** – Build consensus on interoperability and data sharing

http://www.laceproject.eu

This document was produced with funding from the European Commission Seventh Framework Programme as part of the LACE Project, grant number 619424.