Licence management for Public Sector Information

Analysis and modelling of PSI re-use regulation

Vasily Bunakov*, Keith Jeffery**

** Science and Technology Facilities Council, United Kingdom, vasily.bunakov@stfc.ac.uk*

*** Science and Technology Facilities Council, United Kingdom, keith.jeffery@stfc.ac.uk*

Abstract: *The volumes of PSI (Public Sector Information) published on the Internet by national and local governments, international organizations and other public bodies have grown dramatically in recent years. Terms and conditions for this information re-use may differ among suppliers hence need to be analysed and modelled, especially in view of machine assisted or automated processing employed by e-infrastructure and data management projects. The paper presents results of analysis of PSI reuse terms and conditions across several categories of PSI data sources, as well as outlines directions for regulation modelling and its further implementation in software platforms of an infrastructure scale.*

Keywords: PSI (Public Sector Information), license, regulation, analysis, modelling, e-infrastructure

Acknowledgement: This paper is related to the ENGAGE FP7 Project "An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens" (www.engage-project.eu). The authors would like to thank their colleagues of the ENGAGE project for their input for this paper although the views expressed are the views of the authors and not necessarily of the project.

Digital agenda for Europe (ec.europa.eu/digital-agenda/) considers sharing Public Sector Information with citizens and business an important source of sustainable economic growth and knowledge-driven development. PSI is typically thought of as documents issued by State, regional or local authorities, international organizations, other bodies as a result of performing their public duties. This may include economic and demographic indicators, information about environment, healthcare, education and other aspects of a modern society. The proposed revisions of the EC Directive on re-use of public sector information (PSI Directive, 2011) suggest alignment of specific information domains such as scientific information or cultural heritage with PSI domain so it is likely that regulation for all information that is produced or preserved using public funds and under public law will bear more and more similarity as the legislation process progresses.

We discuss challenges for the modelling and implementation of PSI regulation in e-infrastructure platforms based on analysis of national and regional Open Data portals across Europe, with the addition of a few international bodies and remarkable examples from beyond Europe. We then consider modelling techniques and possible design solutions for e-infrastructure platforms in respect to managing PSI regulation, and emphasize the need of a cross-national PSI regulation framework with a technology component in it.

1. Challenges for PSI regulation modelling and implementation

The historical focus on Document rather than Data in the European PSI Directive and some of European Commission Decisions may hinder to a certain extent the development of modern regulation for PSI e-infrastructure but we want to focus on other major challenges identified.

1.1. Amount and structure of regulation

The amount of PSI regulation to be considered for the e-infrastructure design and implementation may seem modest if we take into account only legal statements published on PSI open data sources (PSI portals) as these statements should ideally encapsulate all other regulation so that e-infrastructure could just consider a single document in each case.

The legal statements, however, may refer to underpinning licenses or other regulation, as well as to the exclusions from common terms and conditions; the actual structure of regulation hence adds up to the amount of documentation to be considered. The diagram on Figure 1 shows the structure of the World Bank legal notes with some details omitted, to keep the whole thing readable.

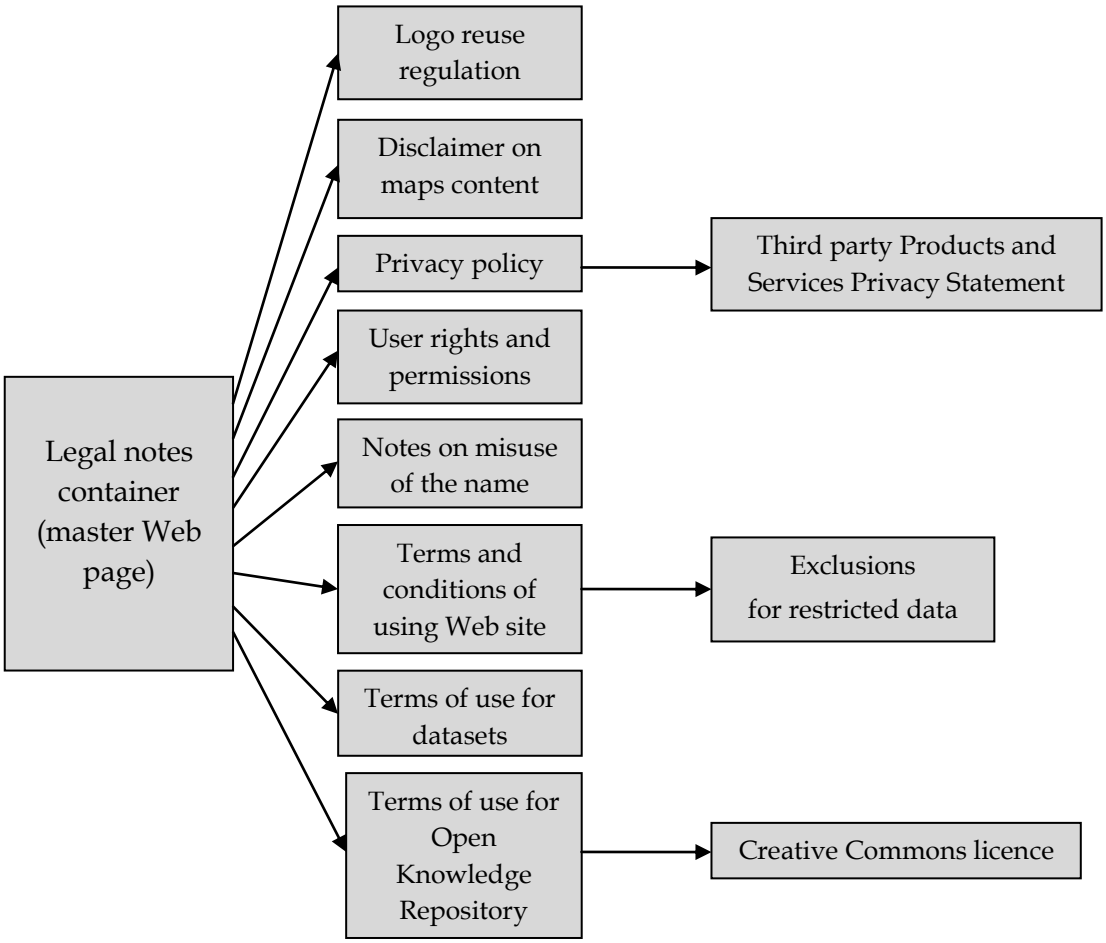

Figure 1: High-level structure of the World Bank legal notes. Each rectangle represents a separate document (Web page).

Despite that the World Bank does not belong to the public sector which is the main focus of our work, any reasonable e-infrastructure for PSI data should incorporate or make linkable the rich data assets of such international bodies so considering their practice of Open Data regulation, and modelling it should be a part of PSI e-infrastructure projects focussed on data re-use. Also the structure of the World Bank Open Data regulation has matured through decades hence can be a sort of a "role model" for relatively recent attempts to formulate the pieces of regulation for PSI portals; it indicates how PSI portals regulation may evolve in years to come.

A particular user of a PSI e-infrastructure platform may not be interested in all categories of legal notes, e.g. her primary concern may be terms of use for datasets but not those for a logo. However, it is in nature of infrastructure projects with e-infrastructure not being an exclusion that one cannot predict the exact modes of infrastructure use, especially in the medium- and long-term; that is why it is important to model the entire structure of regulation associated with data sources that are prominent candidates for data acquisition and data re-use in e-infrastructure.

Another problem is that the metaphor of Document behind a piece of regulation that may well suit human consumers (ideally having a juridical background) may not be adequate for software components of e-infrastructure that need more detailed and interpretable guidance. Hence if we measure the amount of regulation not as the number of different documents encountered but as the number of granular regulation statements in them, it will add up to the volume of regulation to be modelled and processed. The diagram on Figure 2 shows a structure of the information re-use license for the French governmental portal data.gouv.fr The list of components/features may be incomplete and depend on a particular regulation description framework chosen for the granular regulation description; we use Creative Commons categories in this case accompanied by other features worth mentioning for this license.

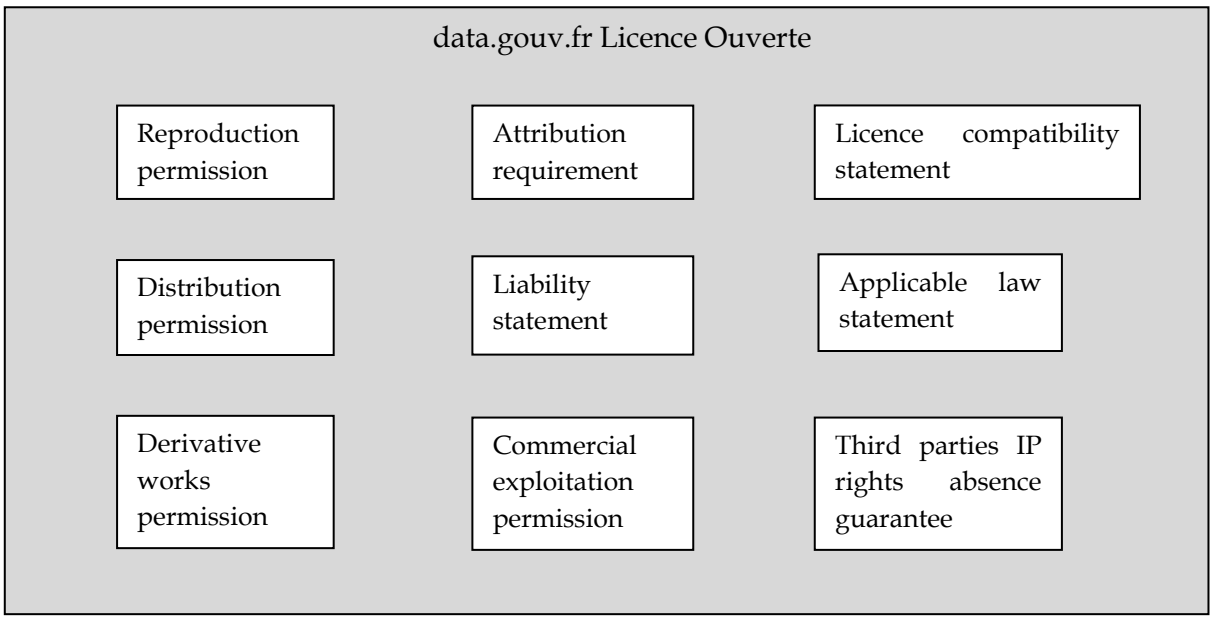

Figure 2: Regulation components of data.gouv.fr open licence. Each white rectangle represents a granular regulation component within the text of the licence (represented by the grey rectangle).

Yet another factor of scale for managing regulation is the granularity of its application: it may be applied to particular data collections within a data source (PSI portal), or to a single dataset. We did not conduct this sort of analysis for PSI data sources but detailed research on data re-use in controlled data collections (many of them being good candidates for linking with PSI data or for ingest in PSI e-infrastructure data stores) shows that up to a half of them offer dataset-level terms of use, and about a third of them – click-through terms of use when one cannot actually reach a dataset via the Web link without having agreed to the terms and conditions (Eschenfelder and Johnson, 2011): these latter ones of course can be generic although nothing prevents them from being specific as the mechanism for the granular publishing of regulation is already there.

1.2. Regulation diversity

Figure 1 gives an idea of typical subjects of regulation in Open Data portals but even for the same subject, regulation may be diverse across data sources of a similar nature like national .gov portals. Our observations on national PSI portals of eight countries show that each of them introduced its own licence for data re-use:

Table 1: Licences of European governmental data portals

Country	Portal	Licence
France	Data.gouv.fr	Licence Ouverte
United Kigdom	Data.gov.uk	Open Government Licence
Italy	Dati.gov.it	Creative Commons Attribuzione - Non commerciale 2.5 Italia (CC BY-NC 2.5)
Germany	Govdata.de	Datenlizenz Deutschland – Namensnennung – Version 1.0 (recommended for common use) Datenlizenz Deutschland – Namensnennung – nicht kommerziell – Version 1.0 (for exceptions)
Norway	Data.norge.no	Norsk lisens for offentlige data (NLOD)
Netherlands	Data.overheid.nl	No specific common licence but a recommendation for the agencies publishing data through the portal to use the framework of the Open Government Act, and to apply Creative Commons Zero of Public Domain if any licence is desired at all
Spain	Datos.gob.es	No specific licence but two parts in extensive legal notes that cover data re-use and are based on different pieces of Spanish national legislation
Belgium	Data.gov.be	No specific common licence. Each public service or government institution determines the terms and conditions governing access to and use of its data published through portal.

This shows that governments take different approaches to licensing their PSI: some of them (France, United Kingdom, Italy, Norway) offer a common licence that covers the portal content by default; Germany offers more than one licence for different modes of data re-use so that the governmental agencies may choose what is more appropriate in a particular case of data publishing; Netherlands provide a certain framework and recommendations but no common licence; some countries (Spain, Belgium) just offer a common data publishing platform where different governmental agencies may apply their own licences.

What is also remarkable is that PSI portals offering a common licence – despite their claims that it is based on open data principles with popular references to Creative Commons – still decided to produce their own flavour of an open licence.

1.3. Regulation updates

Open data licences and other information re-use regulation are possibly not the most frequently updated items yet they are subject to change that has to be managed. We have encountered only one case so far where this issue is taken into account, and only from one specific perspective of how to refer to the newer versions of the licence in case it is changed: this is a French Licence Ouverte that explicitly states that users may keep referencing to the current version of Licence well after its updates.

Specific issues related to the updates may arise because of the chains of regulation where one "child" piece of it is based upon another one, and that basic "parent" item is updated or superseded by a newer legislation. Then the "child" regulation item should be updated after the "parent" was renewed but this is not always the case. An example of this is the European Environment Agency data re-use policy (EEAcopyright, 2012) that states it is based on two particular pieces of legislation with one of them, as our checks showed, having been superseded by a newer act, yet the published policy still bears the reference to an obsolete one. This is the very case when automated or machine-assisted update might help to keep the legal notice current but the unstructured character of it (just a Web page) does not allow any reasonable automation.

1.4. Specific regulation content or structure

Some PSI and open data portals apply specific requirements for their information re-use that may affect the effort required for e-infrastructure platforms to adopt these information sources.

The Basque country open data portal (opendata.euskadi.net) requires granular attribution including the date of the last dataset update. This may add effort required for e-infrastructure implementers to actually satisfy this requirement as it seems to be introduced in view of humans referring to the original source with no specific means for bulk information re-use that should be reasonably automated in order to be efficient.

The Singapore open data portal (data.gov.sg) requires a clear attribution with the suggested exact wording of it. This does not seem to take into account possible updates of this exact wording so someone re-publishing Singapore data may unintentionally breach the licence if the current formula for the reference that is correct at the moment of data citation gets obsolete afterwards; there is no mechanism that would allow re-publisher to stay tuned with the current data citation requirement. Another specific requirement of Singapore open data portal is the necessity for application developers to get registered with the portal; the commercial re-use of the information

also requires registration; these two requirements set certain limits to data acquisition and to the e-infrastructure sustainability models that may require a certain level of commercialization.

The OECD portal (www.oecd.org) imposes some specific requirements that can make its data unhandy for mashing them up with PSI data. Upon re-use, one should cite the title of the material, OECD copyright, publication year (if available) and page number or URL as applicable. Again, this seems to be required with only human consumers in mind but e-infrastructures are likely to employ various software agents for data management; there is currently little or nothing in OECD regulation that appeals to this type of information re-use. Also OECD regulation sets certain limitations for the linking technology that e-infrastructure platform may want to employ, e.g. referencing via Web frames or other visual altering tools is not allowed.

2. Common patterns of PSI regulation

Our analysis suggests not only differences in PSI regulation but some common patterns, too, that provide a valuable input for machine-oriented regulation modelling. We discern between patterns of the regulation content (which means finding commonalities among structural schemas similar to Figure 2) and patterns of its representation, i.e. commonalities for the form in which pieces of regulation are shared.

The important pattern of PSI regulation content is that the information published is typically free for commercial re-use; it is also free of royalties or other charges. This is no surprise as Digital Agenda for Europe and national directives of a similar kind do mean the re-use of PSI to be one of the major drivers behind its publication.

The next important pattern is a requirement of PSI attribution (credits to publisher) when someone re-uses it. The exact formulation of this requirement differs among PSI portals: some of them formulate it in general form, others are more specific up to the requirement of the exact wording that should be placed in any material that refers to the PSI source.

Another common pattern of PSI regulation is that publisher claims no responsibility for the consequences of the information re-use. Some of them specify the very moment when their responsibility becomes void: at the moment when the information leaves their portal, i.e. as soon as someone has it retrieved.

Transformations of the PSI artefacts acquired are typically also allowed, as well as re-dissemination of PSI artefacts unchanged.

A common structural characteristic of many PSI regulation artefacts is referring to national legislation that underpins them. In case of pan-European Open Data sources the role of underpinning regulation is commonly played by EC Directives and Decisions. When modelling this characteristic, it may be worth to introduce a common abstraction that will be instantiated either by national or international legislation.

The remarkable pattern of PSI regulation representation is that published items of it: licences, terms and conditions, legal notes – are always underpinned by the metaphor of Document. The metadata about data shared through PSI portal is often available in a well-structured format but there is no structured metadata for regulation items which are just texts.

Another pattern of representation can be thought of as a placeholder or "a pattern of absence": not only a piece of regulation is a Document, it also does not bear a unique identifier for referencing it. The PSI regulation Documents published can now be referenced only through their

Web addresses which are sometimes remote from being "cool URIs"(CoolURIs, 2008). This is not a merely technical issue as in the absence of permanent identifiers, the information attribution requirement does not have a sustainable model to implement it: someone may supply a reference to the regulation item that tomorrow becomes invalid as the licence issuer has it moved (or even removed), e.g. because the URI naming schema has changed owing to the transition of the entire portal onto a new Web server.

For convenience of their further consideration, we compiled the patterns observed into the table:

Table 2: Common patterns of PSI regulation

Patterns of regulation content and structure	Permission for commercial re-use Permission for information transformation Permission for information re-distribution Requirement of attribution (due credits) Taking no responsibility for information re-use Referring to national legislation
Patterns of regulation representation	Metaphor of Document Absence of unique licence identifier

These patterns and new ones that may emerge later on as a result of systematic monitoring may contribute to the metadata models or profiles of the existing rights management frameworks that will enable machine-assisted semantic sharing of PSI regulation. Some of these models or profiles may be specific for a particular e-infrastructure platform that is targeted at certain user communities; a wider PSI regulation framework that we discuss in the end of this paper will also benefit from further collection and systematization of common regulation patterns.

3. Solutions for PSI regulation modelling and design

The large volumes of text documents encapsulating regulation, their interdependencies, and the need for update consistency all demand the use of ICT (Information and Communication Technologies). We described the aspects of a PSI regulation landscape that appeal to business analysts for application of their techniques to the adequate incorporation of various PSI regulation into emerging e-infrastructures. We now suggest a few particular techniques and approaches that we deem valuable to explore and discuss with information technologists.

3.1. Modelling techniques

In a best case scenario, the human end-user or intelligent software needs to process the regulations as well structured statements with formal syntax and declared semantics. A human can do this from free text (although commonly with misunderstandings); technology is not so smart. Ideally, the regulations would be encoded as first-order-logic rules: IF x THEN y ELSE z, as an example:

IF licence is Creative Commons CC-BY

THEN use the document freely as a human or ICT system with attribution

ELSE next rule

The rules will require persistent identifiers for a piece of regulation as a whole, and for the granular statements in it, e.g. what is "Creative Commons CC-BY" or what is "attribution" should be unambiguously defined. A conventional technique for this is the use of an ontology encoded in description logic and stored either in an extended relational form or as statements in OWL, RDFS, or other knowledge representation language.

Publishing the rules should be ideally combined with publishing a manifest with a reference to a particular metadata model chosen, as well as a reasonable description in terms of this model. There are a few candidate metadata models to choose from: Creative Commons Rights Expression Language (ccREL, 2008), Open Digital Rights Language (ODRL, 2012), an appropriate part of the Asset Description Metadata Schema (ADMS, 2012), eXtensible Rights Markup Language (XrML, 1998), or the rights management extension for METS metadata framework (METSRights, 2005).

The metadata manifest may also incorporate universal metadata frameworks not specifically devoted to the description of rights but essential for semantic interpretation or for effective data sharing. An example of the former is CERIF (Common European Research Information Format)[1] that is very strong in description of organizations and their divisions, as well as the relations of organizations with their outputs (regulation being one of them); an example of the latter is Dublin Core (DC, 2012) supported by popular metadata distribution frameworks such as OAI Protocol for Metadata Harvesting (OAI-PMH, 2012).

This is to show we have a good choice of modelling and design frameworks available; we further discuss some of them in the rest of this paper.

3.2. Possible design solutions

Since the legalistic documents are not – at least now – coded as first order logic we are faced with three possibilities for pre-existing documents:

(1) Just supply the document and let the user determine the usage conditions;

(2) Try to interpret the legalistic document using intelligent software to extract the first order logic;

(3) Consciously design or re-engineer pieces of regulation – licences, terms and conditions – in a structured manner, and supply them with API.

The first option (1) is already useful if the licensing and other legalistic information is (a) attached unambiguously to the document or dataset including linkage to the organisation or person who is the license owner/authorizer and (b) the links have temporal information indicating the period of time during which the link is valid (i.e. the period of time when the organization or person that is the authorizer provides access under the named license).

The second option (2) requires more research although there are research projects indicating some success.[2] Within the ENGAGE project the scope is such that we follow the first option.

For new legalistic documents or for the existing ones reengineered there is a third option (3): to encode the regulations and make it available as metadata. The Dublin Core metadata set (DC, 2012) has limited rights information, also the eXtensible Rights Markup Language (XrML, 1998) is a language designed for such a solution and is standardized as REL (Rights Expression Language)

[1] See under www.eurocris.org

[2] http://docs.marklogic.com/5.0/guide/search-dev/binary-document-metadata

for use with MPEG-21; however this limits its applicability more widely. Creative Commons Rights Expression Language (ccREL, 2008) is associated with Creative Commons and is more applicable; it links properties of the work (document) with properties of the license. However the metadata associated with each is rather limited and while the linkage has some semantics (especially concerning the permitted usage) it lacks the temporal information. An advantage of ccREL is that it is W3C compliant and can be implemented in HTML, XML or RDF.

Asset Description Metadata Schema (ADMS, 2012) is a recently proposed mechanism for describing digital assets and includes the repository holding the asset; the asset, contact information, licence, period of time, publisher, documentation, item, asset type, publisher type, status, license type, representation technique, interoperability, language, theme taxonomy, theme, file format and geographic coverage. However the representation is limited to RDF and less rich expressions.

The third approach (3) of a structured regulation modelling is taken by Linked Content Coalition[3] that is endorsed by the European Commission and some national governments for promotion in media business. This is a good example of a collaborative work by big players in a certain information domain, also an indication of a potential for the machine-oriented modelling and processing in other fields including PSI regulation.

Within ENGAGE we take the first option but leaving open the door to others. In particular we use CERIF (Common European Research Information Format) which has formal syntax and declared semantics and links instances of entities such as persons, organisations, publications (including licenses), products (including datasets) via links with both a role (e.g. permissions) and a temporal interval of validity. Furthermore any entity or attribute may be classified using one or more classification schemes giving great flexibility in cross walking from one scheme to another for interoperability. From CERIF one can generate RDF, XML or HTML and since it provides a richer syntax and semantics than the derivatives it can act as a superset representation.

4. Conclusion: the need of a PSI regulation framework

Our analysis of the actual PSI regulation and its application in governmental data portals shows a diversity of approaches taken, and proves the need of having a common framework that should eventually reconcile the differences; otherwise the regulation may become a barrier to building and exploiting scalable e-infrastructures of a cross-national scale. We consider a few interlinked areas of activity that in our opinion should constitute an infrastructure-oriented PSI regulation framework that will address the issues identified:

- Monitoring and update of national and international legislation: laws, directives, decisions
- Monitoring and update of granular regulation on data re-use: licences, terms and conditions
- Machine-oriented regulation modelling and other information technology
- Standardization and structured communication

[3] www.linkedcontentcoalition.org

All the types of activities in the framework should be interrelated. As an example, the experience of applying a particular licence or terms and conditions may drive the need of a top level legislation update, or technology update. Modelling and IT get input from, and provide feedback to other activities. Standardization and communication through national and international bodies, professional consortiums and alliances allow to share and promote best practices.

This paper has focussed on two components of the PSI regulation framework: analysis of granular regulation on data portal level and the discussion of IT design choices available. A certain emphasis on technology is specifically important because of the co-existence of human users and software agents in any modern e-infrastructure: that is why PSI regulation items associated with data should be well structured, and shared having in view machine or machine-assisted information processing.

Publishing data through public sector portals according to specific practices and tailored regulation well serves the need of public bodies to fulfil their legal obligations and prove their openness in modern ways. This may not be enough, however, for exploiting economic, social, and environmental benefits of data re-use. Further elaboration of the suggested PSI regulation framework should facilitate the effective and efficient data re-use to the benefit of various stakeholders of PSI lifecycle.

References

ADMS (2012). Asset Description Metadata Schema, Version 1.0. https://joinup.ec.europa.eu/asset/adms/release/100

ccREL (2008). Creative Commons Rights Expression Language. http://www.w3.org/Submission/ccREL/

CERIF: Common European Research Information Format. See under www.eurocris.org

CoolURIs (2008). Cool URIs for the Semantic Web. W3C Interest Group Note 03 December 2008. http://www.w3.org/TR/cooluris/

DC (2012). Dublin Core Metadata Element Set, Version 1.1. http://dublincore.org/documents/dces/

EEAcopyright (2012). European Environment Agency data re-use policy. http://www.eea.europa.eu/legal/copyright

Eschenfelder K. and Johnson A. (2011). The Limits of Sharing: Controlled Data Collections. Proceedings of the ASIS&T 2011 annual meeting, New Orleans, October 9-12, 2011. http://mail.asis.org/asist2011/proceedings/submissions/62_FINAL_SUBMISSION.pdf

METSRights 2005. Rights extension for METS (Metadata Encoding and Transmission Standard). http://www.loc.gov/standards/rights/METSRights.xsd

OAI-PMH (2008). The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. http://www.openarchives.org/pmh/

ODRL (2012). Open Digital Rights Language, Version 2.0. http://www.w3.org/community/odrl/two/model/

PSI Directive (2003). Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information.

PSI Directive (2011). Proposed revision of the Directive 2003/98/EC on the re-use of public sector information. See under http://ec.europa.eu/information_society/policy/psi/rules/eu/index_en.htm

XrML (1998). eXtensible Rights Markup Language. http://www.xrml.org/

About the Authors

Vasily Bunakov MSc

Vasily Bunakov is a researcher in the Scientific Computing Department of STFC (Science and Technology Facilities Council). He earlier worked in the IT and research departments of High Energy Physics laboratories in Russia and Switzerland, and for the London branch of Deutsche Bank. His current research interests are focussed on digital preservation and linked open data.

Prof. Dr. Keith Jeffery

Keith Jeffery is currently Director International IT Strategy at STFC (Science and Technology Facilities Council). Keith previously had operational responsibility for IT with 360,000 users, 1100 servers and 140 staff. Keith holds 3 honorary visiting professorships, is a Fellow of the Geological Society of London and the British Computer Society, is a Chartered Engineer and Chartered IT Professional and an Honorary Fellow of the Irish Computer Society. Keith is currently President of ERCIM and past president of euroCRIS, and serves on international expert groups, conference boards and assessment panels.