LSE Research Online

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Dave Puplett

Version identification – a growing problem

Article (Accepted version)

Original citation:
Puplett, Dave (2008) Version identification – a growing problem. *Ariadne* (54).

Link to article: http://www.ariadne.ac.uk/issue54/puplett/

© 2008 UKOLN

This version available at: http://eprints.lse.ac.uk/21496/
Available in LSE Research Online: October 2008

LSE has developed LSE Research Online so that users may access research output of the School. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LSE Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain. You may freely distribute the URL (http://eprints.lse.ac.uk) of the LSE Research Online website.

This document is identical in content to the publisher's version. Should you wish to cite this article please use the publisher's version.

http://eprints.lse.ac.uk

ARIADNE

Version Identification – a growing problem

Dave Puplett outlines the issues associated with versions in institutional repositories, and discusses the solutions being developed by the Version Identification Framework (VIF) project.

Introduction

The problem of version identification in institutional repositories is multifaceted and growing. It affects most types of digital object now being deposited, and will continue to grow if left unaddressed as the proliferation of repositories continues and as they are populated with more and more content.

The JISC (Joint Information Systems Committee) has consequently funded the VIF project as part of the Repositories and Preservation programme, running from June 2007 to May 2008. The project is tasked with developing a framework for the identification of versions of digital objects in repositories and working to secure community acceptance of the recommendations made within the framework.

This paper describes the current situation, informed by the results of two surveys undertaken as part of the project, and previews the solutions likely to be proposed by the project in the form of the framework.

The problem and previous studies

The project team, through study of previous work on the issues around versioning and using the results of the VIF requirements exercise, identified the major issues that repository users and managers have encountered with identifying versions with repositories.

Most significant of previous work in this area are the VERSIONS project (Versions of Eprints – a user Requirements Study and Investigation Of the Need for Standards) [1], and the RIVER (Scoping Study on Repository Version Identification) study [2] which both examined aspects of version identification in the light of the open access movement.

VERSIONS, led by LSE (London School of Economics and Political Science), recently found that 59% of researchers produce four or more types of research output from each research project [3]. Types discussed were articles, book chapters, working papers, conference papers, and presentations amongst others. The VERSIONS project focused upon e-prints in Economics, and found that not only do researchers output these different types of object, but also that each one of these may be developed through several draft versions. These different outputs are increasingly likely to be made available as working papers, work in progress or pre-prints during the development of a piece of research, and VERSIONS found that this was leading to

confusion and time consuming inspection of objects for end users who were trying to identify the work they were finding online.

Previous to the VERSIONS project, the RIVER study concluded that:

> *'The issue of version identification is not simply (indeed not primarily) one of unique identification of resources but rather of defining the relationship between resources. While it is important that each of those resources should be uniquely referenceable, from a user standpoint the more significant questions to ask are*
>
> *• In what way are these two things the same?*
>
> *• In what way are these two things different (and to what extent does that difference matter to me)?'* [4]

The number of potential versions created as part of a contemporary research project is clearly large, and the relationships between iterations, variations and manifestations are theoretically enormous. This is further complicated by projects with co-authorship and collaborative work.

Although the VIF project team has remained wary of wading into the controversial subject of how open access affects the publishing industry, it is clear from both VERSIONS [3] and VIF's own research that a great deal of versioning problems arise from reader confusion over the publication status of the content that they find in a repository.

The significant problem for repository managers is how best to organise these multiple versions within a repository, and how best to describe them so they can be properly found and understood. This problem extends beyond the confines of the single repository implementations and interface to include cross repository searches and other open access harvesters.

Problems identified include:

- Confusion over whether an article is the published version, a copy that is identical in content to this but unformatted, a draft version, an edited version and so on.
- Repository searches often yielding many results of items which ostensibly appear to be the same thing, but actually vary in terms of content, formatting or propriety file type.
- Research outputs with co-authors being deposited in different places at different stages of development without guidance as to which is authoritative or most recent.
- Multimedia items being handled poorly by repositories that treat them as text, and their relationship to other objects that form part of the research project being undefined by the repository.
- Vastly inconsistent approach of different repository software packages and implementations in how versions are dealt with.

The VIF project has moved on in two significant ways from this prior research. Firstly, we have looked at the requirements for a variety of digital objects, not just text documents and across the whole range of disciplines. Secondly, the VIF team has described our approach to versions as agnostic. The research carried out so far prior to VIF has taken a different approach, with much discussion already having taken place on defining and using labels to describe work in the context of the publication process.

VIF therefore defines a version as ‘a digital object (in whatever format) that exists in time and place and has a context within a larger body of work’. This definition allows for ultimately user defined opinions about what constitutes a version, and underpins the project teams’ desire to remain neutral and produce a clear and transparent framework. Our core aim with the framework is to make important information which would facilitate clear version identification to be made available to end users and repository managers.

The results of the VIF user requirements survey

The project undertook two surveys in the Autumn of 2007 of over 100 information professionals working in repositories and 70 academics (mainly in the UK with some international responses).

There are clear results from the surveys, identifying a common feeling amongst repository staff in how great a problem they perceive version identification to be, which versions of research repositories should be ingesting and what version identification solutions would be expected to work best.

The survey highlighted some striking realities about version identification. Important to the success of the framework is the knowledge that approximately a third of Information Professionals with responsibilities toward a repository stated that they either have no system currently in place or ‘don’t know’ how they deal with versioning at present.

There were few instances where opinion was shared between academics and information professionals, but both groups agreed that identification of versions is a problem – only 5% of academics and 6.5% of information professionals surveyed found it easy to identify versions of digital objects within institutional repositories. The situation becomes even worse across multiple repositories (1.8% and 1.1% respectively).

When asked what types of material they currently stored in their repositories, 95.4% of information professionals claimed that they currently store, or plan to store, text documents with many also stating that they store, or plan to store, audio files (73.6%), datasets (77.9%), images (83.3%), learning objects (46.5%) and video files (75.3%).

This awareness of a wider range of object types being created and desire on the part of information professionals to store them is positive, especially in the context of the results of the academics survey, which suggested a large number of researchers either already create or intend to create audio files (47.2%), datasets (68%), images (72.5%), learning objects (74.6%) and video files (57.6%). As expected, the vast majority also intend to continue working with text documents.

Most notable from the survey of academics was the concern that only the most complete version of their work be made available in repositories. This is in direct contrast to the wishes of information professionals, who overwhelmingly wanted to store all versions that a researcher has made available. The academics we surveyed were very clear about their wish to only make the finished version of their output ultimately available. The strength of feeling on this issue was demonstrated by the

comments we received as free text, often even in answers to questions on different subjects.

A corollary of this is the behaviour therefore expected of information professionals by academics when it comes to organising versions of their work within a repository.

The project team realised that as well as addressing practical issues about version identification we will also need to examine issues such as whether repository managers should replace objects when newer versions become available, or retain all versions as representative of the research process. There are subsequent questions about the value of keeping all versions for future research purposes and preservation, and developing policies that suit individual researchers, disciplines, institutions and their repositories.

The project team offered several broad potential solutions to the problem of version identification to all survey respondents. These included chronological and numeric approaches, taxonomies and the use of id tags.

Although support was high for all of the suggestions (only one solution had less than 60% support across the respondents)[5] free text comments revealed numerous and varied misgivings. No one solution stood out as having more support than any other. We encouraged free text responses with each suggested solution, and many qualifications and caveats were given.

The most contentious example is that of using a taxonomy to define the version status. The notion is clearly attractive, and has been addressed previously with NISO / ALPSP (National Information Standards Organisation / Association of Learned and Professional Society Publishers) [6], RIVER and VERSIONS all offering possible vocabularies to describe versions. Many free text comments remarked that whilst the idea is a sound one in principle, implementing such a taxonomy would be virtually impossible without some sort of enforcing body. Also, getting community agreement on the terminology used would be a very controversial subject due to the often polarised standpoints of publishers and information professionals. Serving the best interests of the research itself by setting up a descriptive taxonomy sounds a worthy concept, but insulating it from the pre-established terminology and bias of certain camps would clearly be a very serious undertaking.

This was a blessing and curse for developing solutions for the framework, as none of the suggested solutions had been eliminated and the options left available are complex and not applicable generically. Therefore, we chose to detail many solutions, with their benefits and problems made explicit, and allow the eventual audience of the framework choose the solutions that suit their needs best. The lack of a silver bullet to solve the version identification problem was ultimately no surprise, considering different needs across disciplines and different types of object.

Identifying key stakeholders and developing a suitable framework of solutions

The project identified three major stakeholder audiences to address with the framework. Each has a role to play in facilitating easier version identification in the future.

The most ideal and reliable source of version information about an object must be the author or creator of that object. Awareness of versioning issues amongst content

creators when disseminating their work in repositories is low, and the framework should promote better practice in both embedding version information within an object, and supplying repository staff with the right information when their work is deposited.

Much care must be taken here to keep a light touch and minimise the red tape surrounding repository deposit, but still ensure that sufficient attention is paid to identify an object when it is disseminated in this way. The project is committed to contributing to and improving the repository experience, not imposing obstacles to their success. It is essential therefore for VIF to consider the role of repositories through the eyes of the content creators – what versions are visible etc.

The staff working on actual and future repositories are the key audience for VIF. Repositories across the UK and globally are at early stages of development, and each implementation is different. Different in terms of content stored, with different subjects, object type, collection policies. The staff time available to repositories also varies enormously, from some institutions having full time staff to some having only a few hours here and there. Therefore the framework must be implementable by staff with a range of resources available to them.

The most powerful group the VIF will address are repository software developers, as this group are able to control the exact way versions are organised and represented in repositories (acknowledging of course that much repository software is often open source and customisable by users with the right know-how).

The repository software packages currently available deal with versions in different ways. There are clear pros and cons associated with different systems. For example, Fedora allows the depositor to place new versions on top of old ones, which gives a clear identification of the newest work. However it deals less well with lateral versions, such a Microsoft Word file and a PDF file kept alongside each other. Eprints has a different approach of allowing versions to be kept in separate record that are then linked to each other. However, this only allows for a linear arrangement, as each document can only be either a successor or a predecessor of another.

A further issue concerning the repository software currently available is that even with good intentions, most are still primarily designed, or implemented, to deal with text documents. Work is going on to improve this situation and investigate ways of broadening the use of repositories to deal with different file types better, such as the DataShare project [7], and the Application Profiles for Images [8], Geospatial data [9] and Time based media. These variations will be taken into account by the VIF project team as we try to address future repository software development.

The VIF project team consists of Jenny Brace, LSE, Dave Puplett, LSE, Paul Cave, University of Leeds, and Catherine Jones, Science and Technology Facilities Council. This team has worked in conjunction with an invited expert group, details of which can be found on the VIF Web page: http://www.lse.ac.uk/library/vif, to develop the framework itself and the solutions that it will contain.

A look at how those solutions are taking shape

Having identified the three major stakeholders that the project wishes to address, the project team decided to make the framework targeted to each of these audiences by creating advice and recommendations specific to each group.

Content creators

The first step in the chain of establishing the version status of an item should be to have the item identifying itself. This is fairly straightforward for text documents as they usually contain at least title and author on a front page. This can easily be expanded to include some sort of version information. This could be specific dates, a description, a version number or so on. Such information can also be embedded into the file itself by using the properties fields that most desktop applications employ. The framework gives advice on how to embed version metadata in all sorts of digital object, from text documents written in word or converted to PDF, data in spreadsheets through to images, video and audio created or compressed in a variety of file formats.

If version information is imbued to the document by the creator, it becomes authoritative and available to anybody who then comes into contact with the item, including the repository manager.

It is crucial for version identification that metadata stored for items with multiple and/or ambiguous versions is rich enough to allow for the disambiguation of versions. The project will be making recommendations on ways of achieving this, and what information is important to be made transparent through the repository from the creator to the end user.

One of the messages we will be promoting to content creators is the need to communicate version information as well as traditional bibliographic information when an item is deposited to a repository. This ideally will also include information about anywhere else an item might have been deposited, to begin to address the problem of version identification across multiple repositories. This message must be carefully communicated however, as it is important to minimise the burden upon authors and content creators, which is already seen as a barrier to the success of repositories.

Repository managers

The staff who run repositories on the ground are undoubtedly our most significant audience and the framework is at its greatest depth here. The advice offered is practical, wide-ranging and addresses several different aspects of repository maintenance.

VIF will be recommending that repositories include a statement about versioning in appropriate policy documents, and will make recommendations on what to include in these. We have also made approaches to the OpenDOAR [10] team about improving the support that their template policies offers for versions.

The ingest and metadata capture process of an item is the best opportunity for a repository manager to establish the version status of an item. We will detail ways of

incorporating version awareness into both self deposit systems and proxy deposit undertaken by library staff.

How this information is then incorporated into the metadata for an object, what parts of this information are made transparent to the end user, and what information survives the harvesting process of cross repository search mechanisms are central issues for the framework to deal with.

We will be looking closely at how best to utilise current metadata standards, especially in the light of the recently completed Scholarly Works Application Profile (SWAP) (previously known as Eprints Application Profile) [11]. The use of a FRBRised structure in the SWAP particularly appeals to the VIF project because of the way that metadata is tied more closely, and less ambiguously, to individual 'entities', or objects as I have referred to them here. This application profile allows metadata to describe not only the work in question (such as a particular research project) but the individual versions of this, right down to the level of actual format of each version and distinct copies of it.

The project team will also discuss how to make the best use of other metadata schemes to describe versions, detailing how to use available fields to describe version status, and encourage uniform usage of these to enhance version identification across different repositories. We will also look at recommendations for the future, to raise awareness of version identification as an issue for future metadata development projects.

The framework will also make more minor recommendations that could benefit certain repositories, such as considering implementing a uniform file naming within a repository. Another solution discussed has been the use of cover sheets. Some repositories (for example LSE [12]) insert a cover sheet onto the front of their text documents, which can contain detailed information about the provenance of an item, including version information. The use of coversheets is mainly limited to text documents however.

Software developers

VIF's approach to software developers of repository software will take the form of specifications for future software improvements that would make version identification easier. Part of this task is to improve awareness of the issue amongst this audience that version identification is an important problem.

Some of the specifications have already been implemented by some repository software developers, but the project recommends that the following features are highly desirable across repositories.

The specifications are grouped into two broad categories; those which affect the 'on site' usage of a repository, and those which relate to how metadata is structured within the repository ready to be searched or harvested by third party aggregator.

The first proposed software improvement will be to introduce automatic checking of title and author upon deposit to guard against duplication, and to help make sure

versions are identified as such at this early stage. This feature would prevent an object related to another already held within the repository being deposited with the relationship being flagged up by whatever versioning structure the repository employs.

Ideally this feature would also be able to compare the deposited item to the contents of other repositories and flag up any overlap or relation to other external objects. This would be particularly useful in the case of work by multiple authors working at different institutions. The challenge for software developers is to integrate this into both templates for a self deposit processes or in the back end of the software interface for repository staff.

VIF will also recommend the wider use of thumbnailing to preview objects. This is especially useful for previewing images and multimedia items, where versions are likely to include cropped versions

Much work has already been undertaken on the development of Dublin Core application profiles that provide richer metadata options for repositories to provide for harvesting by open access aggregating services. The framework will recommend that future repository software is designed to support richer metadata export for cross repository searches, such as these application profiles. This could possibly include exporting a thumbnail of an object as part of the metadata, for example.

The project has identified much room for improvement in the way that versions are ranked by third party searches or repositories. The very strong evidence of the VIF survey makes it clear that academics want the latest version of their work to be the one that is found first by end users. Repositories that hold more than one version of an object may well know or record which one supersedes another, but this information is not communicated well externally, leading to erratic ranking results in external searches and confusion over which is the most appropriate version to use.

Summary and conclusions

Open access has made work freely available and opened up many new possibilities for scholarly communication, but managing the way this work is identified and found is essential for repositories to flourish.

If as we all hope, repositories continue to grow, the need for better and clearer version identification will become more and more important as more content is deposited, in different stages of development, in different forms and in different places.

VIF aims to raise awareness of the issues associated with version identification and offer both short term and longer term solutions to the problems identified. Repository managers can do much to improve the situation, with some immediately easy to implement recommendations. Other solutions that we will be making, such as those to software developers, are intended for the longer term.

We will recommend the framework to all three stakeholder groups and encourage it’s advocacy by those who are working in this community to make repositories more accessible, more popular and more successful.

If you are interested in the work of the project, please contact us or look at http://www.lse.ac.uk/library/vif. The framework itself will be officially launched late February 2008 and the project team will be disseminating the framework at events through spring and early summer.

References

1. VERSIONS project http://www.lse.ac.uk/versions
2. Rumsey, S. et al, "Scoping Study on Repository Version Identification (RIVER) Final Report" , Rightscom, 2006 http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf
3. VERSIONS Project Poster http://www.lse.ac.uk/library/versions/VERSIONS_A1_poster_final_(2).pdf
4. Rumsey, S. et al, "Scoping Study on Repository Version Identification (RIVER) Final Report" , 2006, p38 http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf
5. Cave, P. "Work package 2: Requirements exercise - report of a survey of Academics and Information Professionals" 2007, p3 http://www.lse.ac.uk/library/vif/documents.html
6. Recommendations of the NISO/ALPSP Working Group on Versions of Journal Articles, 2006, http://www.niso.org/committees/Journal_versioning/Recommendations_TechnicalWG.pdf
7. DataShare Project http://www.disc-uk.org/datashare.html
8. Images Application Profile http://www.ukoln.ac.uk/repositories/digirep/index/Images_Application_Profile
9. Geospatial Application Profile http://www.ukoln.ac.uk/repositories/digirep/index/Geospatial_Application_Profile
10. OpenDOAR http://www.opendoar.org/tools/en/policies.php
11. SWAP http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile
12. LSE On Line http://eprints.lse.ac.uk/

Acknowledgements

I would like to acknowledge the hard work of Paul Cave, Project Officer for VIF based at University of Leeds for his work of the VIF survey and his subsequent report, from which this article draws results and analysis.

Author Details

Dave Puplett
Project and Communications Officer
London School of Economics and Political Science

Email: d.puplett@lse.ac.uk
Web site: http://www.lse.ac.uk/library/vif

Keywords: please list in order of importance the keywords associated with the content of your article (usually up to 8). These will not be made visible in the text of your article but may be added to the embedded metadata:

1. *VIF*
2. *Versions*
3. *Repositories*
4. *Open-Access*
5. *Metadata*
6. *Self-Deposit*
7. *Research*
8. *Digital*