LIBER QUARTERLY
THE JOURNAL OF THE ASSOCIATION OF EUROPEAN RESEARCH LIBRARIES

Vol. 23, no. 4 (2014) 240–273 | ISSN: 1435-5205 | e-ISSN: 2213-056X

Enhanced Publications: Data Models and Information Systems

Alessia Bardi

Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy and
Dipartimento di Ingegneria dell'Informazione, Università di Pisa, Italy
alessia.bardi@isti.cnr.it

Paolo Manghi

Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, Italy
paolo.manghi@isti.cnr.it

Abstract

"Enhanced publications" are commonly intended as digital publications that consist of a mandatory narrative part (the description of the research conducted) plus related "parts", such as datasets, other publications, images, tables, workflows, devices. The state-of-the-art on information systems for enhanced publications has today reached the point where some kind of common understanding is required, in order to provide the methodology and language for scientists to compare, analyse, or simply discuss the multitude of solutions in the field. In this paper, we thoroughly examined the literature with a two-fold aim: firstly, introducing the terminology required to describe and compare structural and semantic features of existing enhanced publication data models; secondly, proposing a classification of enhanced publication information systems based on their main functional goals.

Igitur publishing | http://liber.library.uu.nl/ | URN:NBN:NL:UI:10-1-116065

Key Words: scholarly communication; data citation; data publishing; executable papers; e-Research; e-Science; data models; information systems

1. Introduction

Scientists are interested in getting scientific reward for their efforts and in learning about research, as well as in being able to re-use the outcome of others' research to reproduce similar experiments and avoid pointless mistakes. To meet such requirements, modern research is being increasingly conducted in so-called e-Science infrastructures and by adopting the collaborative approach of e-Research, which strongly promotes sharing and re-use of all research results, including traditional publications, datasets (Gray, 2007; Lord & Macdonald, 2003; Mooney & Newton, 2012) and "research experiments" (De Roure, Goble, & Stevens, 2009). The uptake of such trends is leveraged by the fact that also funding agencies and organizations, which are crucial stakeholders in the research chain, are advocating and increasingly making mandatory the publishing and citing of any research outcome in order to measure their Return of Investment, improve their funding strategies, or gain visibility and scientific rewards.

The dissemination of research outcomes via traditional publications, either in paper or in digital form, cannot cope alone with these new needs of scientists. To tackle such limits, the new notion of *enhanced publication* has emerged in several scientific disciplines as a new research result dissemination mean. As a traditional digital publication, an enhanced publication is characterised by an *identity* and by *descriptive metadata,* but its nature enables improved and modern ways of implementing dissemination of and access to research material. Existing *enhanced publication information systems* may significantly vary in terms of their functional goals, e.g. Web 2.0 reading and discovery capabilities, and, in recent manifestations, re-production and assessment of scientific experiments. In general, they implement *data models* that describe enhanced publications as consisting of *parts*, whose purpose is to define one mandatory textual description of the research conducted (as for the traditional article) and its relationships with other material whose nature varies depending on the scientific context (e.g. datasets, images, tables, workflows, devices, services) and consumption purposes (e.g. visualization, experiment repetition).

The state-of-the-art on enhanced publications has today reached the point where some kind of common understanding and definition is required, in order

to provide the *terminology* for scientists to *classify*, i.e. comparing, analysing, or simply discussing, the multitude of solutions in the field. A preliminary study in this direction was carried out at SURF Foundation (Woutersen-Windhouwer Brandsma et al., 2009) in the context of the DRIVER-II EC project (Lossau & Peters, 2008). As a result, enhanced publications were defined as "dynamic, versionable, identifiable compound objects combining an electronic publication with embedded or remote research data, extra materials, post publication data, database records, and metadata". The investigation performed an analysis of the requirements for the modelling and management of enhanced publications in a cross-disciplinary scenario and resulted in an abstract and optimal data model, which was used as the basis for a prototype of an enhanced publication information system (Hoogerwerf, 2009; Woutersen-Windhouwer Brandsma et al., 2009). Such work represented an important step in motivating and highlighting the existence of a novel research field and, most importantly, defined the term "enhanced publication" inspiring our study. In this paper, we are willing to extend these initial efforts in two ways:

i. Providing the terminology necessary to describe existing enhanced publication data models in terms of their *structural and semantic features*;
ii. Proposing a classification of existing enhanced publication information systems in terms of their *functional goals*.

Our goal is to help researchers in the field at better discussing and motivating their results, but also newcomers on this subject at organizing and structuring their understanding of the literature.

Paper Outline: Section 2 focuses on data models and identifies special classes of data model parts, which can be used to characterize and compare different enhanced publication data models. Section 3 puts the bias on enhanced publication information systems and classifies existing solutions according to the functional goals they implement, describing for each solution the data model parts they support. Section 4 concludes the paper.

2. Enhanced Publication Data Models

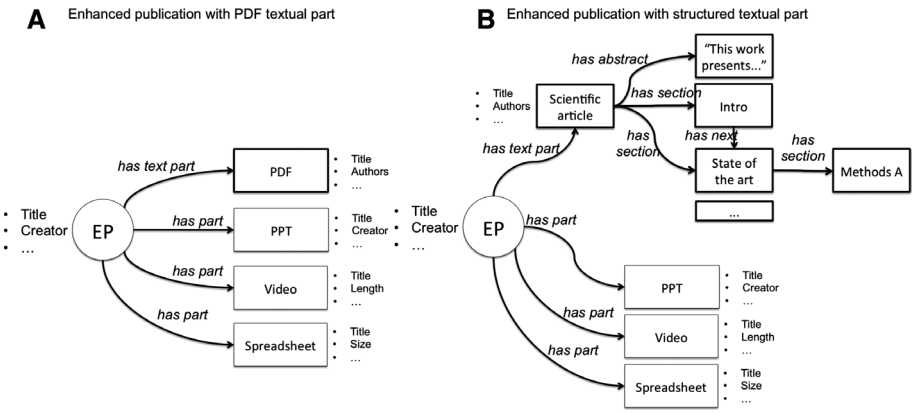
Enhanced publications are digital objects characterized by an identifier (possibly a persistent identifier) and by descriptive metadata information. The

constituent components of an enhanced publication include one mandatory textual narration part (the description of the research) and a set of interconnected sub-parts. Parts may have or not have an identifier and relative metadata descriptions and are connected by semantic relationships. In general, enhanced publication data models vary in the way they define the structure of their parts, metadata, and relationships, which reflect and support the functional goals of a given enhanced publication information system. For example in some approaches an enhanced publication is a "package" embedding all its subparts, i.e. sub-parts cannot be shared by different enhanced publications; in other approaches, parts can instead be referenced, shared or passed as inputs to workflow engines. In some solutions the narrative part of an enhanced publication is intended in a traditional (for the digital world) sense as a readable file (PDF, DOCX, etc.); while in others, the text is structured into interconnected sub-parts, e.g. sections, figures, tables. Figure 1 shows two enhanced publications consisting of the narrative part, representing the scientific article, and supplementary material: the slides presented at a conference, the video of the presentation, and one spreadsheet of related research data. The two publications have an analogous structure for the supplementary material, but differ in the nature of the mandatory narrative: a single PDF file versus a "structured text", i.e. the text is made of several interconnected sub-parts such as abstract, sections, figures, tables, etc. The first model is generally preferable in digital library settings where the traditional management of PDF articles is to be enriched with supplementary material. The second model is more adequate for information systems supporting UIs for the advanced reading experience of the articles, e.g. for navigating or searching through article subparts.

In our analysis we studied the literature of enhanced publication data models trying to identify repetitive patterns of their constituent parts, relationships between them, and associated metadata. The aim was to produce a terminological framework for describing and comparing such data models based on their structural and semantic features.

For metadata descriptions of parts we observed that they provide information at different interpretation levels, so as to enable both human and machine interpretation. Examples are bibliographic information, file information, provenance information, visualization information, execution information, versioning information, etc. In fact, each part may bear several metadata descriptions in order to allow for multiple usages. Relationships between

Fig. 1: Examples of enhanced publications.

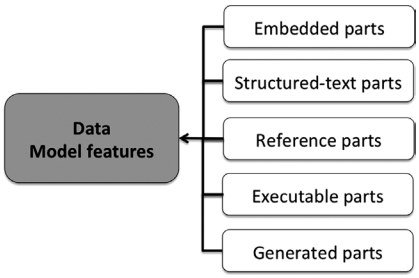


parts implicitly or explicitly (e.g. by a label) characterise the semantics of the association between two parts. Associations may indicate user-oriented links, e.g. *chapterOf, relatedWith, datasetUsed*, application-oriented links, e.g. *alternativeVisualization, externalLink, localLink*, and others typologies. Such classifications have been produced before (Candela et al., 2008) and are common to other scholarly communication fields, such as traditional digital library systems, digital archives, scientific data repositories, or scientific communication infrastructures, e.g. OpenAIRE infrastructure (Manghi, Bolikowski, Manola, Schirrwagen, & Smith, 2012), Swedish ScienceNet (Johansson & Ottosson, 2012), and CRIS systems (Jörg, 2010). As such they may be regarded when describing the specificities of individual data models, but they are not able to capture the peculiar nature of enhanced publications data models. On the other hand, our investigation revealed that the constituent sub-parts of enhanced publications, independently from the metadata descriptions and relationships accompanying them, are the most characterizing aspect of enhanced publication data models.

More specifically, after dissecting existing data models, we observed that the following classes of parts were recurring (Figure 2):

- *Embedded parts*, e.g. enhancing a publication with supplementary material files;
- *Structured-text parts*, e.g. enhancing a publication providing an editorial structure of its textual sub-components;

Fig. 2: Enhanced Publications data model features.



- *Reference parts*, e.g. enhancing a publication with URLs to external objects;
- *Executable parts*, e.g. enhancing a publication with parts that include software and data to run an experiment.
- *Generated parts*, e.g. enhancing a publication with tables which may dynamically change depending on updates of given input research data.

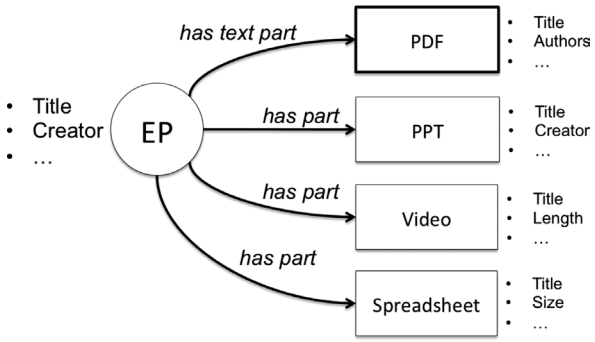
In the next sections we shall describe such classes, motivating their nature and exemplifying their occurrence.

2.1. Embedded parts

A real-world example of data models with embedded parts is that of those publishers that intercepted the need of researchers to add "context" to scientific publications in order to improve the comprehension of their results. To this aim, they provide information systems where authors can upload so-called *supplementary material* along with the publication. Examples are presentation slides, appendices to the article, data and description of data used for the research described in the article, high resolution images, tables that could not be inserted fully in the digital publication because of page limits.

More generally, data models with embedded parts describe enhanced publications whose parts may be files, generally not described by metadata and without an identifier, hence not searchable or sharable by different publications. Typically, embedded parts are stored locally in the information system. The semantics of the relationships between the publication's narrative part

Fig. 3: Enhanced publication with embedded parts.



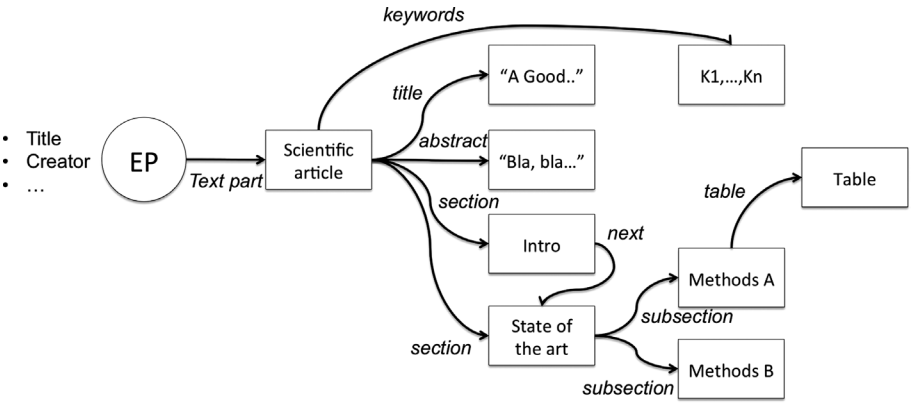
and supplementary material is often "silent", although in some cases it may bear information about the type or the meaning of the files (Candela, Castelli, & Pagano, 2009a; Kircz, 2002). Figure 3 shows an example of an enhanced publication with embedded parts, where each part is accompanied by metadata descriptions.

2.2. Structured-text parts

Some modern approaches abandon the notion of textual publication as a single block of text, e.g. a PDF file, and experiment with the definition of publications as structured texts. Such solutions are often addressed by publishers to enhance publication readability via web 2.0 applications or given client applications. For example, Elsevier proposed the "Article of the Future" (Aalbersberg, Heeman, Koers, & Zudilova-Seinstra, 2012), which implements a Web-oriented viewer of a scientific publication, where the reader can browse through abstract, sections, paragraphs, view and download tables and images, generate PDF manifestations, interact with tables, etc.

Data models with structured-text parts describe enhanced publications with a narration part that is a structured object composed of several interconnected parts, such as abstract, sections, figures, tables, bibliography, etc. As shown in Figure 4, these data models precisely define the relationships linking sub-parts and may include metadata describing the role of the sub-parts (e.g. chapter, section, table) in order to support advanced visualization and reading functionalities.

Fig. 4: Enhanced publication with structured-text parts.



Interestingly, the advent of structured-text enhanced publication information systems opened new challenges on how to minimize the effort required to construct such publications. Typically, the structured-text parts of the publication are identified by humans supported by processing techniques; e.g. full-text mining of PDF publications (Mathiak, Kupfer, Münch, Täubner, & Eckstein, 2005; Peng & McCallum, 2004; Ramakrishnan, Patnia, Hovy, & Burns, 2012; Shah, Perez-Iratxeta, Bork, & Andrade, 2003), XSLT processing of XML publications (Hoeppner, 2013).

2.3. Reference parts

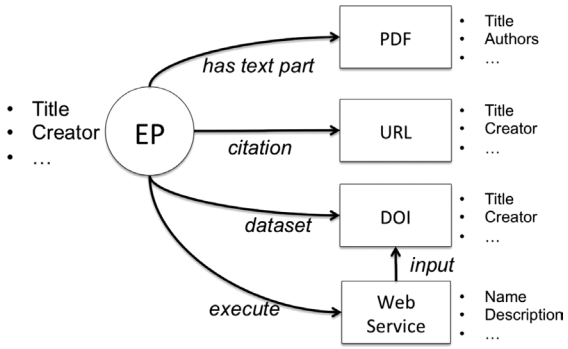
In modern scholarly communication a fully "embedded parts" approach is generally not sustainable and effective. Firstly, the information system at hand should be able to store and handle all parts required for contextualizing a publication and disseminating it as an enhanced publication. For example, publication and dataset management (e.g. storage, preservation, description) are typically complex and separate activities (Castelli, Manghi & Thanos, 2013), handled by professionals through specialized tools. Secondly, in many cases the objects to be referred by the enhanced publications already exist, i.e. they are in fact re-used, and are stored into "external" information systems. Finally, embedded parts do not promote sharing and re-use since enhanced publication sub-parts are reachable and re-usable only via the mandatory narrative part, i.e. they do not have identity and metadata. These limitations are overcome by enhanced publications that feature links to remote

research outputs, such as datasets, other (enhanced) publications (e.g. citations), or supplementary material (e.g. web sites, presentations). For example *data papers* (Newman & Corke, 2009) are the result of the recent trend to give scientific value to the production of datasets and reward their creators by publishing datasets as peculiar research outcomes. Data papers are enhanced publications with a narrative part and a persistent reference part to the dataset stored in a remote data repository. Similarly, in many communities, literature reporting on experimental results is often referring to datasets used or generated by the experiment and vice versa. Other examples of reference parts are those supported by information systems capable of mining publication text to extract URL pointers to scientific knowledge (Lee, Lee, & Kim, 2001; Nayak, Witt, & Tonev, 2002), concepts represented in general-public Web sources [e.g. Wikipedia (Giles, 2005; Haigh, 2011; Page, 2011; Reavley et al., 2012), DBPedia[1]], or discipline-specific databases, ontologies, and taxonomies (e.g. WoRMS[2], UNIPROT[3]).

Figure 5 shows an example of an enhanced publication with reference parts. Data models with reference parts describe enhanced publications whose parts may be references to objects that are "external" to the enhanced publication, hence are possibly shared with others. Such references may be community-specific identifiers (e.g. UKPubMed identifiers), unique persistent identifiers (e.g. DOIs), or URLs. The simplest implementation consists of a traditional PDF article with relative metadata that together become an enhanced publication with reference parts thanks to the addition of dedicated metadata fields containing references to external objects. This solution is generally low cost as it marries the traditional file-metadata approach of Digital Libraries, but referenced parts can be discovered only via the publication metadata. In other systems, reference parts are explicit enhanced publication sub-parts enriched with metadata information inherited (somehow collected) from the referred object, thereby enabling the proper consumption (e.g. discovery, visualization) of the part. For the same purposes, in some cases the semantics of relationships between sub-parts is provided [e.g. SURF enhanced publications (Woutersen-Windhouwer Brandsma et al., 2009)].

For enhanced publications with reference parts the supporting information systems might have to face the issues of "broken links". This is the case when the objects identified by such references are no longer available for any reason. As a result, the enhanced publications may become inconsistent and recovery or invalidating policies may be applied.

Fig. 5: Enhanced publication with reference parts.



2.4. Executable parts

Data-intensive e-Science brought in the novel requirements of disseminating traditional digital publications with a "research experiment context", which would allow for better interpretation and validation by-repetition of the research conducted. Enhanced publications with reference parts to datasets certainly represent a step ahead in this direction, but are not sufficient to address these needs. Indeed, in order to share an experiment scientists should make available to the community both the data and the processes they used. To this aim, information systems capable of managing and consuming enhanced publications with executable parts have been realized. Such parts carry the information required to execute a process, such as a reference to a web service used in an experiment, a workflow to be executed by a given engine, or, more generally, code that can be dynamically executed by a given run-time.

The most important studies on enhancing publications with executable parts for the purposes of supporting peer-review, research validation and re-usability have been conducted in the context of virtual research environments (VREs). VREs are defined by JISC as digital, distributed platforms that enable "collaboration between researchers and provide access to data, tools and services through a technical framework that accesses a wider research infrastructure" (Carusi & Reimer, 2010). VREs offering functionalities for the planning, execution and sharing of in-silico experiments are also referred to as e-laboratories. Examples of e-laboratories are myExperiment (De Roure,

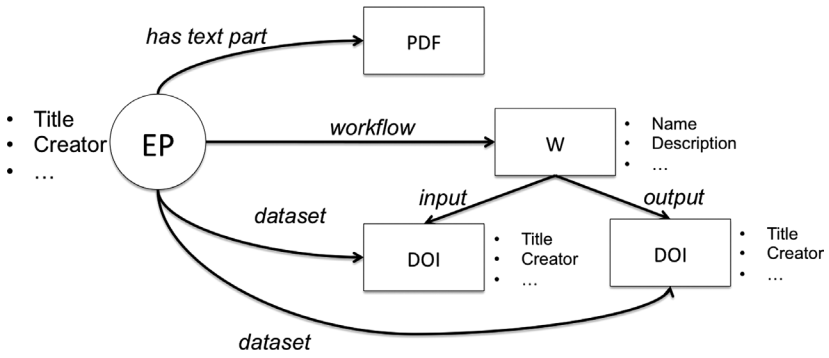
Goble, & Stevens, 2009), Collage (Nowakowski et al., 2011), IODA (Siciarek & Wiszniewski, 2011), SHARE (Van Gorp & Mazanek, 2011), D4Science (Candela, Castelli, Pagano, & Simi, 2005).

Figure 6 sketches an enhanced publication about an experiment, described by a traditional PDF publication relative to a workflow W that generated a given output dataset from a given input dataset. To this aim, the enhanced publication includes all parts required to repeat the experiment: the PDF, the workflow, the two datasets (or a link to their location) and their metadata information. The latter convey information required to execute the workflows and services in the proper way (e.g. the workflow engine, operating system configuration). Researchers, but also reviewers, are therefore in the condition of reproducing the experiment and comparing the results with those presented in the article, thus validating the research results. In addition, they can also apply the workflow W to their own datasets, hence effectively re-use tools produced by others.

Executable parts may be encoded in a variety of ways, which highly depend on the scenario supported by the information system that generates, stores and offers access to such enhanced publications:

- The system hosts the processes themselves: parts point to an executable process (e.g. web service);
- The system relies on third-party execution environments: parts contain code to be executed [e.g. workflows in myexperiments.org (De Roure et al., 2009)];

Fig. 6: Enhanced publications with executable parts.



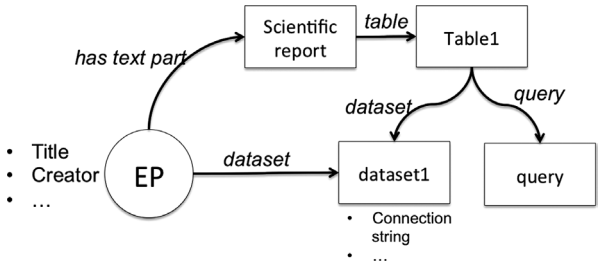
- The system relies on the ability of researchers to install and deploy software: parts include software (e.g. SVN references), how-to manuals, or configurations (e.g. virtual machine configuration).

2.5. Generated parts

In Elsevier's Article of the Future (Aalbersberg et al., 2012) an enhanced publication contains reference parts that link to external databases, e.g. molecule entries in a chemical database. Such parts can be processed by an application of the information system so as to dynamically generate new parts of the enhanced publication. For example, a molecule can be processed to generate its 3D rendering. As such, the rendering is not a static sub-part of the publication, but rather a dynamically generated part. Another example of generated parts is that of the enhanced publication in Figure 7, inspired by the "live documents" proposed by Candela et al. (2008). The publication includes a "scientific report" and a pointer to a database dataset, whose content is constantly updated. The scientific report has a dynamic table whose content can be generated by running a given query over the dataset. When users visualize the scientific report, a local application executes the query and processes the results to generate the table with the content currently available in the dataset.

Generated parts can be defined in terms of (i) the "static" parts of the enhanced publication used as inputs (e.g. the reference to the molecule, the reference to the database dataset, the query); and (ii) the application used by the information system to consume the existing part (e.g. an application to generate 3D models of molecules, a software capable of sending the query to

Fig. 7: Enhanced publication with generated parts.



the database and generate the table from the result). All the parameters and configuration needed to dynamically create a generated part are either in the status of the information system or stored with the metadata of the already existing parts. As a consequence, generated parts are tightly coupled with the information system at hand, with all the re-use limitations that such an approach may entail.

3. Enhanced Publication Information Systems

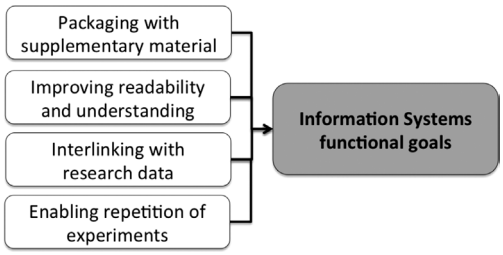
The main motivations behind enhanced publications are to be found in the limits of traditional scientific literature to describe the whole context and outcome of a research activity. The goal is to move beyond the simple PDF (*FORCE11*)[4] to support scientists with digital and automated access to the literature and any form of research outcome (e.g. research datasets, ontologies), still without losing the narrative spirit of "the publication" as a dissemination mean.

A wide variety of information systems for enhanced publications are available out there, which provide communities of scientists with selections of functionalities handling enhanced publications conforming to the most variegated data models. More specifically, such systems offer a set of *management functionalities* for the creation, deletion and updates of enhanced publications, and a set of *consumption functionalities* for the usage of these publications, e.g. reading, sharing, executing.

As to management functionalities, enhanced publications are typically created (deleted and updated) via end-user interfaces that guide end-users towards providing publication parts and specifying relationships between them. Depending on the data model, parts may be provided by uploading files from the file system or providing references to files (e.g. URLs, DOIs). As shown in Section 2.2, in some cases sub-parts may be identified or generated at run-time by the applications. In other cases they may be produced by systems (e.g. scientific infrastructures, virtual environments) for example to provide the machine-executable parts necessary to share a repeatable experiment.

Consumption functionalities, being the driving requirements and motivation for this field, require more attention. Their focus varies and may include

Fig. 8: Enhanced publications information systems: functionality goals.



for example: exporting enhanced publications as packages of parts (i.e. "compound objects"), improving readability of the narrative text via web interfaces or clients (e.g. navigating and visualizing subparts), browsing/accessing parts following relationships (e.g. links from publications to datasets), machine execution of parts (e.g. execution of a dataset processing workflow), etc. This section shapes up a "historical evolution" of the approaches in terms of the scientific motivations and the functional advances that brought us from digital publications on our desktops, e.g. PDFs, to today's executable publications in e-Science infrastructures. In our itinerary we identified four main scientific motivations: *packaging with supplementary material*, *improving readability and understanding*, *interlinking with research data*, and *enabling repetition of experiments*. For each of these areas (Figure 8) we report on relevant solutions in the literature, sorting them in chronological order, and reporting the data model features they implement.

3.1. Packaging with supplementary material

The first enhancement introduced to move beyond the mere digitization of the publication and to investigate new avenues in the digital scholarly communication was likely accompanying a digital publication (e.g. PDF file) with supplementary material. In such scenarios, scientists can share packages consisting of publication and supplementary material in an attempt to better deliver hypothesis and results of the research presented in the publication. In this section we report on enhanced publication information systems whose main functional goal is that of handling packaging a publication with supplementary material. Table 1 lists the information systems in this category and highlights the features of their data models.

Table 1: Provision of supplementary material: information systems and data model features.

Information System	Embedded parts	Structured-text parts	Reference parts	Generated parts	Executable parts
Journals with supplementary material policies	√	Depends on the journal	Depends on the journal	Depends on the journal	
Modular article	√	√	√		
LORE			√		

Several scientific journals (e.g. publishers Elsevier[5], SAGE journals[6]) offer authors the possibility to upload with their article any relevant material that is too big or that does not fit the traditional article format or its narrative, e.g. datasets, multimedia files, large tables, animations, code, etc. Scientists accessing the article online may also download such material to complete their understanding. Supplementary material is typically stored locally into the information system of the journal and it is not discoverable and not accessible outside the context of the related article, i.e. users can find and access the material only after they have discovered and accessed the article. Information systems of journals willing to offer this functionality usually adopt data models where supplementary materials are treated as embedded parts of the enhanced publication.

Kircz (2002) proposes the "modular article" data model. A modular article aggregates and connects with meaningful links several "modules" in order to create a coherent, self-contained, and complete unit describing a research. The modular structure facilitates re-use and re-purposing of modules and enables the realization of web-oriented publishing/reading tools. A module describes, with a specific set of metadata, an entity related to a research. The proposed modules are: meta-information (i.e. bibliographic info, structure of the article, classification terms, references, acknowledgements, abstract), goal and setting (i.e. problem definition, methods, techniques, and goals), results (i.e. raw and fitted data), discussion, and conclusions. Modules may describe/include different types of entities, which reflect the kinds of supplementary materials usually submitted along with a publication, such as: sounds, videos, data sets, and images. The modular article data model is therefore richer than the one proposed by scientific journals, since it features embedded, reference, and structured-text parts.

LORE (Gerber & Hunter, 2010) is a plugin for Mozilla Firefox that allows creating "compound objects". Compound objects are here intended as enhanced publications composed of multiple Web resources and related bibliographic records. The main goal of LORE is that of supporting eLearning in humanities. Teachers and researchers are provided with an easy to use tool to create packages of related resources. Those packages are easy to export and share on the Web, helping students in finding interesting self-contained resources about a specific topic or a course. Resources composing a compound object are related by typed relationships defined in a dedicated and configurable OWL ontology. Since LORE's goal is that of enabling the "packaging" of already existing Web resources, its data model only features reference parts.

3.2. Improving readability and understanding

This category of approaches focuses on enhanced publication data models whose parts, metadata and relationships are defined with the purpose of improving the end-user experience when visualizing and discovering research materials. These approaches are the natural extension of the traditional publication, oriented to reading, and typically integrate all tools made available by the web infrastructure and its data sources (Jankowski, Scharnhorst, Tatum, & Tatum, 2013). Specifically, they explore the possibilities of: (*i*) structuring narrative text into interconnected sub-parts, (*ii*) re-using the universe of web resources to enrich the text, and (*iii*) including dynamic forms of content within the text. Table 2 lists the information systems in this category and highlights the features of their data models.

The D4Science infrastructure (Candela, Castelli & Pagano, 2009b) used in the iMarine project[7] serving fishery scientists, implements the notion of *live documents* introduced by Candela et al. (2005). Live documents consist of textual publications (typically research reports) that embed data descriptions, tables, histograms, summaries, and statistics based on "live data", generated at access time and updated in the publication by the underlying infrastructure. A publication can therefore be "instantiated" in a given moment in time to describe current status/results for a given scenario. The data model of live documents features embedded, reference, structured-text, and generated parts.

SciVee (Fink & Bourne, 2007) is a system that allows authors to create enhanced publications by uploading an article they have already published

Table 2: Improved reading experience: Information systems and data model features.

Information System	Embedded parts	Structured-text parts	Reference parts	Generated parts	Executable parts
D4Science	√	√	√	√	
SciVee	√	√			
PLOS Neglected Tropical Diseases	√	√	√	√	
Veteran Tapes project	√	√			
Utopia documents	√	√	√	√	
Rich Internet Publications	√	√	√		
SOLE		√	√	√	
Article of the future	√	√	√	√	
Bookshelf	√	√	√		

and a video or podcast presentation that describes the highlights of the paper. The author can synchronize the video with the content of the article (text, figures, etc.) such that the relevant parts of the article appear as the author discusses them during the video presentation. Videos and podcast presentations are represented in the data model as embedded parts of the enhanced publications. The SciVee data model also features a "light version" of structured-text parts: the images are extracted from the article text and available in a dedicated viewer of the Web User Interface. However, the system does not process the narrative part of the enhanced publication to identify sections, references, etc.: the narrative textual part is still visualized as a single block of text and the system does not offer browsing and searching functionalities.

Shotton, Portwin, Klyne, and Miles (2009) experiment different enhancements for research articles published in *PLoS Neglected Tropical Diseases* in order to improve the user reading experience. Articles are enriched with semantic tags that identify keywords (e.g. names of diseases, organisms, proteins) and interesting entities such as institutions, dates, and persons. Moreover, interactive tables, figures, and maps are also provided. The features supported by the data model are embedded, reference, structured-text, and generated parts.

In Humanities, van den Heuvel, van Horik, Sanders, Scagliola, and Witkamp (2010) experiment on enhanced publications in the context of the Veteran Tapes project. Authors asked researchers to create manually some samples of enhanced publications. Researchers were provided with a web application to select sound fragments of interviews to link to phrases of their publication.

The resulting enhanced publication is shown to the reader as a traditional digital publication, but parts of the narrative text are associated to audio fragments and highlighted for easy recognition. The reader can click on those parts to view the metadata and the transliteration of the audio fragment. An audio player is also provided for listening to that specific fragment of the interview. The data model of the Veteran Tapes project allows including audio fragments as embedded parts of enhanced publications. Structured-text parts are also supported, but limited to the parts that are associated to audio fragments: other elements of the article are not identified and exploited to provide other kinds of advanced reading facilities such as navigation through sections.

Utopia Documents (Attwood et al., 2010) is a desktop PDF reader that integrates visualization and data-analysis tools with published research articles. Utopia Documents has been used by editors of the *Biochemical Journal* to curate preprints by annotating terms with links to online resources such as Wikipedia, UniProt entries, and ontologies. By exploiting the content of supplementary material attached to a publication and content of related datasets, the software is also able to provide dynamic views of tables and images. Hence, the data model supports the following features: embedded, structured-text, reference, and generated parts.

Further on, Breure, Voorbij, and Hoogerwerf (2011) present the concept of "Rich Internet Publications" (RIPs). A RIP is an online publication composed of narrative text, images, videos, links to data and other material. Text has not a major role with regards to the other components: all parts are "first-class citizens" in RIP's data model. The importance of one specific part of a RIP just depends on the actual reader, who might prefer to start his or her reading from an image, rather than from the description of the data. Since all components are "peers", users have the possibility to choose non-linear (i.e. not guided by the text) reading. Breure et al. support their interpretation by providing several real-case examples in the cultural heritage domain. The data model features embedded, structured-text, and reference parts.

SOLE (Pham, Malik, Foster, Di Lauro, & Montella, 2012) is a tool for linking research papers with associated *science objects*, such as source codes, datasets, annotations, workflows, packages, and virtual machine images. Authors of SOLE are investigating the possibility of enabling re-use of datasets linked by a SOLE document via given services. The features supported by the data model of SOLE are: structured-text parts, reference parts, and generated parts.

Elsevier presents the so-called "Article of the Future" (Aalbersberg et al., 2012), where a publication is structured in well-defined parts (e.g. sections, figures, tables, images, links to external data or applications) to provide end-users with advanced visualization functionalities such as interactive charts and tables, access to geospatial data on Google Maps, 3D fossil reconstruction and chemical compound viewers. The underlying technology mines the narrative parts to identify chemical and science material compound terms (keys to external databases) and assigns them to web services that display detailed information or graphical representations. The data model therefore features embedded, reference, structured-text, and generated parts.

Bookshelf (Hoeppner, 2013) is a system for life science and health care literature resources managed by the National Center for Biotechnology Information (NCBI). Documents in Bookshelf are represented in XML format and are tagged by curators with references to several molecular databases such as OMIM and GenBank. Publications are enriched with embedded supplementary files, structured-text parts that enable advanced navigation of the publications, and reference parts representing the links to molecular database entries added by curators.

3.3. Interlinking with research data

Scientific communities, organizations, and funding agencies are nowadays supporting and welcoming initiatives, standards and best practices for publishing and citing on the Web datasets and publications (Mooney & Newton, 2012; Reilly, Schallier, Schrimpf, Smit, & Wilkinson, 2011). Examples are DataCite[8], EPIC and CrossRef, which establish common best practices to assign metadata information and persistent identifiers to datasets and publications. Data publishing and citation practices are today strongly advocated by research communities, which need datasets to become first citizens in the research production chain. Datasets should be discoverable, reusable, and scientists who produce them should be scientifically rewarded for their efforts (Callaghan et al., 2012; Parsons & Fox, 2013).

Today the trend is depositing data (Piwowar, Vision, & Whitlock, 2011) in discipline specific data repositories — e.g. Pangaea (Diepenbroek et al., 2002) and DataOne (Michener et al., 2011) for earth and environmental science, Archaeology Data Service for archaeology[9], DRYAD for sciences and

medicine (White, Carrier, Thompson, Greenberg, & Scherle, 2008) — or databases (e.g. UniProt[10] and ArrayExpress[11] for biomedical science). Several solutions are investigating the possibility to identify relationships between publications and datasets either prior or post publishing. In the first case, authors include in the publication text or in the metadata the (persistent) identifiers (e.g. URLs, DOIs) of the related datasets or database entries. This is often obtained by reciprocal agreements between scientific journals, which mandatorily require dataset archiving practices (e.g. DRYAD JDAP[12]), and the aforementioned data repositories. In the second case, links between publications and research data are identified after the publishing step, either by humans or machinery. For example, full-text mining techniques can be applied to the narrative part of an enhanced publication in order to identify and attach to the publication the links to datasets (i.e. as corresponding reference parts).

In this section we report on enhanced publication information systems whose main functional goal is that of offering the possibility to enrich a publication with links to relevant research data, in order to strengthen data citation, facilitate data re-use, and reward the precious work underlying data management procedures. Table 3 lists the information systems in this category and highlights the features of their data models. It is worth noticing that several information systems presented in the previous sections also perform this kind of enhancement, but they were not included in this section since the creation of links between articles and data is not their main functional goal.

SCOPE (A Scientific Compound Object Publishing and Editing System) (Cheung, Hunter, Lashtabeg, & Drennan, 2008) is a system designed to enable

Table 3: Publication-dataset linking: Information systems and data model classes.

Information System	Embedded parts	Structured text parts	Reference parts	Generated parts	Executable parts
SCOPE		√	√		
CENS ORE aggregations			√		
EuropePMC	√	√	√		
WormBase			√		
DANS Enhanced Publication Project			√		
BioTea		√	√	√	
OpenAIRE multi-disciplinary EPs			√		

scientists to easily author, publish and edit scientific "compound objects". These are enhanced publications encapsulating (or referring to) datasets and resources generated or utilized during a scientific experiment or discovery process. Such objects, whose structure is represented as an RDF "named" graph, are then published and exchanged [also via OAI-ORE (Lagoze et al., 2012)] to disseminate the overall context of a research activity. The SCOPE data model supports structured-text and reference parts features.

Pepe, Mayernik, Borgman, and Van de Sompel (2009) present an implementation of OAI-ORE (Lagoze et al., 2012) to aggregate publications, datasets, and metadata related to the seismology and environmental sciences at the Center for Embedded Networked Sensing (CENS). Entities are linked to each other via relationships whose semantics belong to the scholarly and scientific life cycle. Since the goal of the system is to link resources already published on the Web, the only supported feature of the data model is that of reference parts.

EuropePMC, formerly known as UKPMC (McEntyre et al., 2011), is an information system that aggregates life science abstracts and Open Access full-text publications from several sources. Publications are enriched with references to entries in well-known biomedical databases (e.g. UniProt, ENA, PDB), ontologies and taxonomies (e.g. NCBI Taxonomy). Relevant entries are identified by applying text-mining techniques to the full-texts (Rebholz-Schuhmann, Arregui, Gaudan, Kirsch, & Jimeno, 2008). The data model of EuropePMC[13] supports enhanced publications that consist of (i) one structured-text narrative part composed of abstract, sections, and bibliography; (ii) embedded supplementary material, as it was originally delivered by the authors to the publisher; and (iii) a list of references to database, taxonomy and ontology entries.

WormBase (Yook et al., 2012) is a model organism database that integrates research literature, genomic sequences and other biology related aspects. WormBase applies text-mining techniques to research articles to identify keywords in the published literature and links WormBase entries to the relevant publications. Collaborations with journals are currently active in order to support "back-links", from literature to WormBase entries, and increase the discoverability of both database entries and publications. The data model of WormBase represents links to publications as reference parts, including metadata of the original publication.

GreyNet[14] (Grey Literature Network Service) and DANS[15] (Data Archiving and Networked Services) co-operates for the Enhanced Publications Project (Farace, Frantzen, Stock, Sesink, & Rabina, 2012). The goal of the project is to enhance GreyNet's collections of conference pre-prints with links to the underlying research data. Full-text available on GreyNet can be linked to research data archived in the online archiving system of DANS or any other public data archive. DANS ensures to maintain bi-directional links in order to support discoverability of research data (full-texts in GreyNet link to DANS entries) and of literature (research data stored in DANS link to the related full-texts in GreyNet). The only supported feature of the data model is, at the time of writing, that of reference parts.

The Biotea platform (Garcia Castro, McLaughlin, & Garcia, 2013) aims at integrating scientific literature with the Web of Data. Biotea transforms XML encoded articles into the RDF format. RDF files are then enriched with machine-generated annotations about: (i) the structure of the article (e.g. sections, paragraphs.); (ii) references to biological databases, and (iii) names of entities corresponding to well known ontologies. Biotea features a prototype GUI for the visualization of documents where enriched content is displayed in an "interactive zone" (e.g. molecule viewer, additional information about a protein). In summary, Biotea implements a data model that features structured-text parts, reference parts targeting biological database and ontologies, and generated parts (used by the interactive zone of the GUI).

As part of the OpenAIRE scholarly communication infrastructure, Hoogerwerf et al. (2013) have built demonstrators of multi-disciplinary enhanced publication information systems in the fields of Social Sciences, Humanities, and Life Sciences. Enhanced publications are created by researchers who select a to-be-enhanced publication from the OpenAIRE infrastructure information space (an aggregation of publication metadata from European institutional repositories) and identify links to datasets in several disciplines via semi-automated user interfaces. The enhanced publication is assigned a new identifier and relative metadata (the author is the researcher who created the enhanced publication), and can be searched for and visualized by other researchers. As such, enhanced publication parts, including the narrative part, are encoded as reference parts enriched with metadata from the original objects.

3.4. Enabling repetition of experiments

How can research results be effectively "peer-reviewed" or more generally evaluated for their quality? The creation of enhanced publications is certainly an important step ahead in this direction, since researchers can count on further information contextual to the experimentation described in the research paper. However, even the acquisition of research data used in the experimentation is often not sufficient in many research fields to assess quality of results. Indeed, scientists and reviewers should be equipped with the tools necessary to repeat the experiment, and the authors should share such tools as part of their enhanced publication. In fact, this does not sound as a surprise in e-Science infrastructures, where researchers are urging for tools to disseminate and re-use the whole context of their research. To address such demands, several solutions were proposed in the literature on how to share executable components via enhanced publications. In enhanced publications with executable parts, narrative parts are accompanied by executable workflows with the purpose of reproducing experiments. Table 4 lists the information systems in this category and highlights the features of their data models.

Hunter (2006a) describes the concept of Scientific Model Packages (SMPs), and its evolution in Scientific Publication Packages (Hunter, 2006b). SMPs are defined as compound objects that embed and relate heterogeneous components such as data, software, workflows, graphs, tables, or publications. Hunter also presents a VRE where scientists can create a workspace to share all materials about a research investigation. Scientists can then define SMPs by selecting components in the workspace. Different SMPs might be generated for different purposes (e.g. one for e-learning, one to support peer-review).

Table 4: Repetition of experiments: information systems and data model classes.

Information System	Embedded parts	Structured text parts	Reference parts	Generated parts	Executable parts
Scientific Model Packages	√	√	√		√
myexperiment.org	√		√		√
IODA	√	√	√	√	√
Paper Mâché	√		√		√
Collage Authoring Environment	√	√	√	√	√
SHARE	√		√		√

An implementation where SMPs are implemented and exported in RDF is also available (Hunter & Cheung, 2007). The model of SMPs supports embedded, structured-text, reference, and executable parts.

In 2007, the project *myExperiment.org*, co-funded by JISC and Microsoft, was opened to the public. The goal of the project is to enable scientists to create, share, re-use and re-purpose workflows for data analysis. The earlier work on the design of the VRE (De Roure et al., 2009) led to the definition of the "Research Objects" data model (Bechhofer, De Roure, Gamble, Goble, & Buchan, 2010). A Research Object is defined as a "unit of knowledge" that aggregates resources related to a scientific experiment or a research investigation. Resources include publications, bibliographic metadata, results of an experiment, the data used and produced (or a link to it), methods applied to produce or analyse the data, and the people involved in the research. Named relationships link resources belonging to the same Research Object. The names of the relationships express the semantics of the associations in the context of the scientific investigation. The features supported by the Research Object data model are: embedded, reference, and executable parts.

IODA (Interactive Open Document Architecture) (Siciarek & Wiszniewski, 2011) is an XML-based data architecture for the creation of "Executable Digital Objects" starting from existing electronic publications. IODA creates three layers over the publication: the data layer, the information layer, and the knowledge layer. The data layer identifies the structural entities of the article (sections, images, references, etc.) and enables the association of the paper with embedded or remote research data and executable code. The information layer allows authors to specify links between the structural entities and executable code (e.g. re-generate the plot by running this code when the user clicks on that figure). The knowledge layer allows authors to define links between parts of different Executable Digital Objects, thus enabling the construction of graphs of related Executable Digital Objects. The data layer and the information layer of IODA's data model implement the embedded, reference, structured-text, generated and executable parts features.

The tool Paper Mâché (Brammer, Crosby, Matthews, & Williams, 2011) proposes a method based on virtual machines that provide an environment for authoring, reviewing, and publishing executable papers. Virtual machines, which include all the required tools and software setup for the

reproduction of the experiment, are uploaded into the system with the publication. The virtual machine may also contain data (or a link to it), the required scripts and embedded code snippets to generate updated revisions of a paper and allow reviewers to trace back the steps and verify the results of the authors. The data model features embedded, reference, and executable parts.

Nowakowski et al. (2011) won the Executable Paper Grand Challenge[16] organised by Elsevier with the Collage Authoring Environment. Collage is a VRE where authors can create and share with reviewers "Executable Papers". Authors can enhance their publication, thus transforming it into an Executable Paper, by embedding or linking to the primary research data and executable code. Executable code attached to a publication may be added to (i) enable readers/reviewers to access primary data; (ii) enable interactive generation of figures and plots; (iii) allow the reader/reviewer to execute a computation. The data model adopted by Collage thus features embedded, text-structured, reference, generated and executable parts.

In 2011, the second prize of the Executable Paper Grand Challenge went to SHARE (Sharing Hosted Autonomous Research Environments) (Van Gorp & Mazanek, 2011). The main point of SHARE is that whenever a publication is about a new algorithm or software, the traditional publication is peer-reviewed, but the algorithm and software are not because reviewers might have non-compatible Operating Systems, insufficient hardware requirements, etc. SHARE tries to overcome these issues by supporting researchers in the creation of remote Virtual Machine Images (VMIs) where all software and data related to a publication can be deployed and configured. The resulting VMI is therefore an environment where reviewers can validate and evaluate the paper results without having to download or install anything on their local computers. The data model of SHARE features embedded parts, reference parts and executable parts.

4. Conclusions

In modern science, scientific communities and funding agencies recognize the importance of sharing research results together with their experimental context. Given such objectives, the traditional publication paradigm reveals

several limits and the notion of "enhanced publication" has been extensively investigated and developed to fill these gaps. As a result, the literature offers today a plethora of information systems specifically devised to manage enhanced publications that address the scholarly communication requirements of a given application domain. Examples of publication "enhancements" are: packaging publications with other research material; improving reading experience via Web or desktop clients; creating relationships between publications and research datasets; and ultimately, repeating scientific experiments described by publications.

In this paper we aimed at introducing common terminology and classification schemes in order to shed some light and put some order in such a rich but foundationless realm. More specifically, we identified common structural features of data models and then classified information systems in terms of their main functional goals. As preliminary output of this survey, two considerations can be made.

Existing information systems tend to be delivered adopting ad-hoc technical implementations. Their feature is that of thoroughly satisfying the needs of the final users for whom they were conceived, but in general they fail to be re-used in different contexts, where data model and functional requirements may slightly or heavily change. In fact, being in its early stage of factorization and foundation, this research field lacks general-purpose technical solutions, in the direction of Enhanced Publication Management Systems. Such systems would enable developers to easily realize and maintain enhanced publication information systems starting from their data models.

Enhanced publications are certainly worth considering as proved by the strong interest and investment of scientific communities. Their wide adoption and usage is today hindered by their discipline specific character, which calls for radically different data models and information systems for different communities, and by the higher manual efforts required to draft enhanced publications, which require authors to approach a learning curve without any certainty of getting scientific benefits. For such reasons, today enhanced publications are more frequent in those contexts where their creation and access are imposed by given policies. Examples are scientific journals enforcing enhanced publication formats, e-Science infrastructures enforcing procedures to publish publications together with reproducible experiments. Due to the strong motivations behind research data sharing and the valuable impact

of research infrastructures, which provide communities with common services for experimental reproduction, we hope and expect these difficulties to dissolve in the future, both at discipline and cross-discipline levels.

Acknowledgements

This work is partially supported by the European Commission as part of the project OpenAIREplus (FP7-INFRA-2011-2, Grant Agreement no. 283595).

References

Aalbersberg, I. J., Heeman, F., Koers, H., & Zudilova-Seinstra, E. (2012). Elsevier's article of the future enhancing the user experience and integrating data through applications. *Insights: The UKSG Journal, 25*(1), 33–43. doi: 10.1629/2048-7754.25.1.33.

Attwood, T. K., Kell, D. B., McDermott, P., Marsh, J., Pettifer, S. R., & Thorne, D. (2010). Utopia documents: linking scholarly literature with research data. *Bioinformatics, 26*(18), i568–i574. doi: 10.1093/bioinformatics/btq383. Retrieved May 7, 2013, from http://bioinformatics.oxfordjournals.org/content/26/18/i568.full.pdf+html.

Bechhofer, S., De Roure, D., Gamble, M., Goble, C., & Buchan, I. (2010). *Research objects: Towards exchange and reuse of digital knowledge*. Retrieved May 7, 2013, from http://imageweb.zoo.ox.ac.uk/pub/2010/Proceedings/FWCS2010/05/Paper5.pdf.

Brammer, G. R., Crosby, R. W., Matthews, S. J., & Williams, T. L. (2011). Paper mâché: Creating dynamic reproducible science. *Procedia Computer Science, 4,* 658–667. doi:10.1016/j.procs.2011.04.069. Retrieved May 7, 2013, from http://ac.els-cdn.com/S187705091100127X/1-s2.0-S187705091100127X-main.pdf?_tid=f67bd05a-b758-11e2-ae67-00000aacb35d&acdnat=1367960529_17442f5f9384c4e8a45715b21ab31e0d.

Breure, L., Voorbij, H., & Hoogerwerf, M. (2011). Rich internet publications: "Show what you tell". *Journal of Digital Information, 12*(1). Retrieved May 7, 2013, from http://journals.tdl.org/jodi/index.php/jodi/article/view/1606/1738.

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., et al. (2012). Making data a first class scientific output: Data citation and publication by NERC's Environmental Data Centres. *International Journal of Digital Curation, 7*(1), 107–113. doi: 10.2218/ijdc.v7i1.218. Retrieved May 7, 2013, from http://www.ijdc.net/index.php/ijdc/article/view/208/277.

Candela, L., Castelli, D., & Pagano, P. (2009a). OpenDLib: A digital library service system. In Y. Theng, S. Foo, D. Goh & J. Na (Eds.), *Handbook of research on digital libraries: Design, development, and impact* (pp. 1–7). Hershey, PA: Information Science Reference. doi: 10.4018/978-1-59904-879-6.ch001.

Candela, L., Castelli, D., & Pagano, P. (2009b). D4Science: an e-Infrastructure for Supporting Virtual Research Environments. In M. Agosti, F. Esposito, C. Thanos (Eds.), Post-proceedings of the Fifth Italian Research Conference on Digital Libraries – IRCDL 2009 (pp. 166–169). Delos, Network for excellence on digital libraries.

Candela, L., Castelli, D., Pagano, P., & Simi, M. (2005). From heterogeneous information spaces to virtual documents. In E. A. Fox & E. J. Neuhold (Eds.), *Digital libraries: Implementing strategies and sharing experiences* (pp. 11–22). Berlin, Heidelberg: Springer.

Candela, L., Castelli, D., Ferro, N., Ioannidis, Y., Koutrika, G., Meghini, C., Pagano, P., Ross, S., Soergel, D., Agosti, M., Dobreva, M., Katifori, V., & Schuldt, H. (2008). *The DELOS Digital Library Reference Model – Foundations for Digital Libraries (Version 0.98)*. 2007 from http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_098.pdf.

Carusi, A., & Reimer, T. (2010). *Virtual research environment collaborative landscape study*. JISC, Bristol. Retrieved May 7, 2013, from http://www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf.

Castelli, D., Manghi, P., & Thanos, C. (2013). A vision towards Scientific Communication Infrastructures. *International Journal on Digital Libraries*, *13*(3–4), 155–169. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/s00799-013-0106-7.

Cheung, K., Hunter, J., Lashtabeg, A., & Drennan, J. (2008). SCOPE: a scientific compound object publishing and editing system. *International Journal of Digital Curation, 3*(2), 4–18. doi: 10.2218/ijdc.v3i2.55. Retrieved May 7, 2013, from http://www.ijdc.net/index.php/ijdc/article/view/84/55.

De Roure, D., Goble, C., & Stevens, R. (2009). The design and realisation of the Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems 25*(5), 561–567.

Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., & Wefer, G. (2002). PANGAEA – an information system for environmental sciences. *Computers & Geosciences, 28*(10), 1201–1210.

Farace, D. J., Frantzen, J., Stock, C., Sesink, L., & Rabina, D. L. (2012). Linking full-text grey literature to underlying research and post-publication data: An Enhanced Publications Project 2011–2012. *The Grey Journal, 8*(3), 181–189. Retrieved February 19, 2013, from http://www.opengrey.eu/data/70/01/53/GL13_Farace_et_al_2012_Conference_Preprint.pdf.

Fink, J. L., & Bourne, P. E. (2007). Reinventing scholarly communication for the electronic age. *CTWatch Quarterly, 3*(3), 26–31. Retrieved May 7, 2013, from http://www.ctwatch.org/quarterly/articles/2007/08/reinventing-scholarly-communication-for-the-electronic-age/.

Garcia Castro, J. L., McLaughlin, C., & Garcia, A. (2013). Biotea: RDFizing PubMed Central in support for the paper as an interface to the Web of Data. *Journal of Biomedical Semantics, 4*(Suppl 1), S5. Retrieved October 4, 2013, from http://www.jbiomedsem.com/content/4/S1/S5.

Gerber, A., & Hunter, J. (2010). Authoring, editing and visualizing compound objects for literary scholarship. *Journal of Digital Information*, *11*(1). Retrieved May 7, 2013, from http://journals.tdl.org/jodi/index.php/jodi/article/view/755.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438*(7070), 900–901. doi: 10.1038/438900a.

Gray, J. (2007). A transformed scientific method. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The Fourth Paradigm: Data Intensive Scientific Discovery*. Redmond, WA: Microsoft. Retrieved May 7, 2013, from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf.

Haigh, C. A. (2011). Wikipedia as an evidence source for nursing and healthcare students. *Nurse Education Today, 31*(2), 135–139. Retrieved May 7, 2013, from http://download.journals.elsevierhealth.com/pdfs/journals/0260-6917/PIIS0260691710000912.pdf.

Hoeppner, M. A. (2013). NCBI Bookshelf: books and documents in life sciences and health care. *Nucleic Acids Research, 41*(D1), D1251–D1260. doi:10.1093/nar/gks1279. Retrieved May 7, 2013, from http://nar.oxfordjournals.org/content/41/D1/D1251.full.pdf+html.

Hoogerwerf, M. (2009). Durable enhanced publications. *Proceedings of African Digital Scholarship & Curation*. Retrieved May 7, 2013, from http://www.ais.up.ac.za/digi/docs/hoogerwerf_paper.pdf.

Hoogerwerf, M., Lösch, M., Schirrwagen, J., Callaghan, S., Manghi, P., Iatropoulou, K., et al. (2013). Linking and enriching data and publications across subject-specific infrastructures – Challenges and issues for a multidisciplinary approach. In *Proceedings of the 8th International Digital Curation Conference*, Amsterdam. PowerPoint retrieved May 7, 2013, from http://www.dcc.ac.uk/sites/default/files/documents/IDCC13presentations/Hoogerwerf2IDCC13.pdf.

Hunter, J. (2006a). *Scientific models – A user-oriented approach to the integration of scientific data and digital libraries.* Retrieved May 7, 2013, from http://www.valaconf.org.au/vala2006/papers2006/55_Hunter_Final.pdf.

Hunter, J. (2006b). Scientific publication packages – A selective approach to the communication and archival of scientific output. *International Journal of Digital*

Curation 1, 33–52. doi: 10.2218/ijdc.v1i1.4. Retrieved May 7, 2013, from http://www.ijdc.net/index.php/ijdc/article/view/8/4.

Hunter, J., & Cheung, K. (2007). Provenance Explorer – a graphical interface for constructing scientific publication packages from provenance trails. *International Journal on Digital Libraries, 7*, 99–107. doi: 10.1007/s00799-007-0018-5.

Jankowski, N. W., Scharnhorst, A., Tatum, C., & Tatum, Z. (2013). Enhancing scholarly publications: Developing hybrid monographs in the humanities and social sciences (January 10, 2012). *Scholarly and Research Communication, 4*(1): 010138, 26 pp. Retrieved May 7, 2013, from http://src-online.ca/index.php/src/article/viewFile/40/123.

Johansson, Å., & Ottosson, M. O. (2012). A national Current Research Information System for Sweden. In *e-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production*, Uppsala University, University Administration, pp. 67–71, Agentura Action M.

Jörg, B. (2010). CERIF: The common European research information format model. *Data Science Journal*, *9*, CRIS24–CRIS31. Retrieved May 7, 2013, from https://www.jstage.jst.go.jp/article/dsj/9/0/9_CRIS4/_pdf.

Kircz, J. G. (2002). New practices for electronic publishing 2: New forms of the scientific paper. *Learned Publishing, 15*, 27–32. doi: http://dx.doi.org/10.1087/095315102753303652. Retrieved May 7, 2013, from http://docserver.ingentaconnect.com/deliver/connect/alpsp/09531513/v15n1/s4.pdf?expires=1368021284&id=74113678&titleid=885&accname=Guest+User&checksum=D96C2BE551BCA2AF16B92BA63C967104.

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S., Sanderson, R., & Johnston, P. (2012). A Web-based resource model for scholarship 2.0: object reuse & exchange. *Concurrency and Computation: Practice and Experience, 24*(18), 2221–2240. doi: 10.1002/cpe.1594.

Lee, J-W., Lee, K. H., & Kim, W. (2001). Preparations for semantics-based XML mining. In N. Cercone, T.Y. Lin, & X. Wu (Eds.), *ICDM, Proceedings IEEE International Conference on Data Mining* (pp. 345–352). IEEE Computer Society. doi: 10.1109/ICDM.2001.989538.

Lord, P., & Macdonald, A. (2003). *e-Science curation report*, prepared for the JISC Committee for the Support of Research. The Digital Archiving Consultancy Ltd. Retrieved May 7, 2013, from http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf.

Lossau, N., & Peters, D. (2008). DRIVER: Building a sustainable infrastructure of European scientific repositories. *Liber Quarterly 18*(3/4), 437–448. Retrieved May 7, 2013, from http://liber.library.uu.nl/index.php/lq/article/view/7942/8215x.

Manghi, P., Bolikowski, L., Manola, N., Schirrwagen, J., & Smith, T. (2012). OpenAIREplus: the European Scholarly Communication Data Infrastructure. *D-Lib*

Magazine, *18*(9–10). Retrieved May 7, 2013, from http://www.dlib.org/dlib/september12/manghi/09manghi.html.

Mathiak, B., Kupfer, A., Münch, R., Täubner, C., & Eckstein, S. (2005). Mining PDF documents for pictures. In B. Berendt, A. Hoth, et al. (Eds.), *European Web Mining Forum* (EWMF 2005) (pp. 52–63).

McEntyre, J. R., Ananiadou, S., Andrews, S., Black, W. J., Boulderstone, R., Buttery, P., et al. (2011). UKPMC: a full text article resource for the life sciences. *Nucleic Acids Research, 39*(suppl 1), D58–D65. doi: 10.1093/nar/gkq1063. Retrieved May 7, 2013, from http://nar.oxfordjournals.org/content/39/suppl_1/D58.full.pdf+html.

Michener, W., Vieglais, D., Vision, T., Kunze, J., Cruse, P., & Janée, G. (2011). DataONE: Data observation network for earth-preserving data and enabling innovation in the biological and environmental sciences. *D-Lib Magazine, 17*(1), 3. Retrieved May 7, 2013, from http://www.dlib.org/dlib/january11/michener/01michener.html.

Mooney, H., & Newton, M. P. (2012). The anatomy of a data citation: Discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication, 1*(1), eP1035. doi: 10.7710/2162-3309.1035 . Retrieved May 7, 2013, from http://jlsc-pub.org/jlsc/vol1/iss1/6/.

Nayak, R., Witt, R., & Tonev, A. (2002). Data mining and XML documents. In *International Conference on Internet Computing, IC'2002*, June 24–27, 2002, Las Vegas, Nevada (pp. 660–666). Retrieved May 7, 2013, from http://eprints.qut.edu.au/1476/1/1476.pdf.

Newman P., & Corke P. (2009). Editorial: Data papers – Peer reviewed publication of high quality data sets. *International Journal of Robotic Research, 28*(5), 587–587. doi: 10.1177/0278364909104283. Retrieved May 7, 2013, from http://ijr.sagepub.com/content/28/5/587.full.pdf+html.

Nowakowski, P., Ciepiela, E., Harężlak, D., Kocot, J., Kasztelnik, M., Bartyński, T.. et al. (2011). The Collage Authoring Environment. *Procedia Computer Science, 4*, 608–617. doi: 0.1016/j.procs.2011.04.064.

Page, R. D. M. (2011). Linking NCBI to Wikipedia: a wiki-based approach. *PLoS Currents, 3,* RRN1228. doi: 10.1371/currents.RRN1228. Retrieved May 7, 2013, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3080707/.

Parsons, M. A., & Fox, P. A. (2013). Is data publication the right metaphor? *Data Science Journal, 12*(0), WDS32–WDS46. doi: 10.2481/dsj.wds-042.

Peng, F., & McCallum, A. (2004). *Accurate information extraction from research papers using conditional random fields. HLT-NAACL*, 329–336. doi: 10.1016/j.ipm.2005.09.002. Retrieved May 7, 2013, from http://people.cs.umass.edu/~mccallum/papers/hlt2004.pdf.

Pepe, A., Mayernik, M. S., Borgman, C. L., & Van de Sompel, H. (2009). *Technology to represent scientific practice: Data, life cycles, and value chains.* CoRR abs/0906.2549. Retrieved May 7, 2013, from http://arxiv.org/vc/arxiv/papers/0906/0906.2549v1.pdf.

Pham, Q., Malik, T., Foster, I., Di Lauro, R., & Montella, R. (2012). SOLE: Linking research papers with science objects. In P. Groth & J. Frew (Eds.), *Provenance and annotation of data and processes, Lecture Notes in Computer Sciences 7525* (pp. 203–208). Springer.

Piwowar, H. A., Vision, T. J., & Whitlock, M. C. (2011). Data archiving is a good investment. *Nature, 473*, 285–285. doi: 10.1038/473285a.

Ramakrishnan, C., Patnia, A., Hovy, E., & Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. *Source Code for Biology and Medicine*, *7*(1), 7. doi: 10.1186/1751-0473-7-7. Retrieved May 7, 2013, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3441580/pdf/1751-0473-7-7.pdf.

Reavley, N. J., Mackinnon, A. J., Morgan, A. J., Alvarez-Jimenez, M., Hetrick, S. E., Killackey, E., et al. (2012). Quality of information sources about mental disorders: a comparison of Wikipedia with centrally controlled web and printed sources. *Psychological Medicine, 42*(8), 1753. doi:10.1017/S003329171100287X.

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., & Jimeno, A. (2008). Text processing through Web services: calling Whatizit. *Bioinformatics, 24*(2), 296–298. doi: 10.1093/bioinformatics/btm557. Retrieved May 7, 2013, from http://bioinformatics.oxfordjournals.org/content/24/2/296.full.pdf+html.

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on integration of data and publications.* Opportunities for Data Exchange (ODE). Retrieved May 7, 2013, from http://www.libereurope.eu/sites/default/files/ODE-ReportOnIntegrationOfDataAndPublication.pdf.

Shah, P. K., Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2003). Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics, 4*(5), 20. Retrieved May 7, 2013, from http://www.biomedcentral.com/content/pdf/1471-2105-4-20.pdf.

Shotton, D., Portwin, K., Klyne, G., & Miles, A. (2009). Adventures in semantic publishing: Exemplar semantic enhancements of a research article. *PLoS Computational Biology, 5*(4), e1000361. doi: 10.1371/journal.pcbi.1000361. Retrieved May 7, 2013, from http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000361.

Siciarek, J., & Wiszniewski, B. (2011). IODA – an interactive open document architecture. *Procedia Computer Science, 4*, 668–677. doi: 10.1016/j.procs.2011.04.070. Retrieved May 7, 2013, from http://ac.els-cdn.com/S1877050911001281/1-s2.0-S1877050911001281-main.pdf?_tid=34df85d8-b805-11e2-8a83-00000aacb35d&acdnat=1368034507_bf84eb1f0a5f0bb1c9bbf0e388806439.

Van den Heuvel, H., van Horik, R., Sanders, E., Scagliola, S., & Witkamp, P. (2010). The VeteranTapes: Research corpus, fragment processing tool, and enhanced publications for the e-Humanities. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)* (pp. 2687–2692). Retrieved May 7, 2013, from http://repository.ubn.ru.nl/bitstream/2066/85921/1/85921.pdf.

Van Gorp, P., & Mazanek, S. (2011). SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science, 4*, 589–597. doi: 10.1016/j.procs.2011.04.062. Retrieved May 7, 2013, from http://ac.els-cdn.com/S1877050911001207/1-s2.0-S1877050911001207-main.pdf?_tid=62c13b9c-b808-11e2-9311-00000aacb362&acdnat=1368035873_1c1a42f1ca58d6ae2f5470ae66e5b21f.

White, H. C., Carrier, C., Thompson, A., Greenberg, J., & Scherle, R. (2008). The Dryad data repository: a Singapore framework metadata architecture in a DSpace environment. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications* (DCMI '08) (pp. 157–162). Dublin Core Metadata Initiative. Retrieved May 7, 2013, from http://dcpapers.dublincore.org/pubs/article/view/928.

Woutersen-Windhouwer Brandsma, R., Verhaar, P., Hogenaar, A., Hoogerwerf, M., Doorenbosch, P., Derr, E., Ludwig, J., Schmidt, B., & Sierman, B. (2009). *Enhanced Publications: Linking Publications and Research Data in Digital Repositories*. Amsterdam: Amsterdam University Press. Retrieved December 17, 2013, from http://dare.uva.nl/document/150723.

Yook, K., Harris, T. W., Bieri, T., Cabunoc, A., Chan, J., Chen, W. J., et al. (2012). WormBase 2012: more genomes, more data, new website. *Nucleic Acids Research, 40*(D1), D735–D741. doi: 10.1093/nar/gkr954. Retrieved May 7, 2013, from http://nar.oxfordjournals.org/content/40/D1/D735.full.pdf+html.

Notes

[1] *DBPedi:* http://dbpedia.org/About.

[2] *WoRMS ontology:* http://www.marinespecies.org/.

[3] *Uniprot databas:* http://www.uniprot.org/.

[4] *FORCE11*: http://www.force11.org.

[5] *Elsevier Supplementary Data policies*: http://www.elsevier.com/journals/vaccine/0264-410X/guide-for-authors#87000.

[6] *SAGE Journals, Author Guide to Supplementary Files*: http://www.uk.sagepub.com/repository/binaries/doc/Supplemental_data_on_sjo_guidelines_for_authors.doc.

[7] *iMarine, Data e-Infrastructures Initiative for Fisheries Management and Conservation of Marine Living Resources*: http://www.i-marine.eu.

[8] DataCite: http://www.datacite.org/.

[9] *Archaeology Data Service*: http://archaeologydataservice.ac.uk/.

[10] *Uniprot database:* http://www.uniprot.org/.

[11] *Array Express database:* http://www.ebi.ac.uk/arrayexpress/.

[12] *DRYAD repository:* http://datadryad.org/pages/jdap.

[13] *EuropePMC:* http://europepmc.org/.

[14] *GreyNet:* http://www.greynet.org.

[15] DANS-KNAW*:* http://www.dans.knaw.nl/en.

[16] *Executable Paper Grand Challenge*: http://www.executablepapers.com/.