

Repositories and the Semantic Web

Pascal-Nicolas Becker | Technische Universität Berlin | Open Science Days | Berlin, October 13, 2014

Agenda

- Repositories
- Semantic Web and Linked Data
- Data exchange with repositories
- Extending repositories
- DSpace 5

xxx.lanl.org / ArXiv.org

“Although the WorldWideWeb still represents only a small fraction of the overall usage, this access mode is expected to become dominant in the near future.”

Paul Ginsparg 1994

Source: Paul Ginsparg, *First Steps Towards Electronic Research Communication*. In: *Computer in Physics*, Vol. 8, No. 4, 1994, pp. 390-396.

Digital Repositories

Repositories are systems to safely store and publish digital objects and their descriptive metadata.

Not in the meaning of software repositories.

Examples:

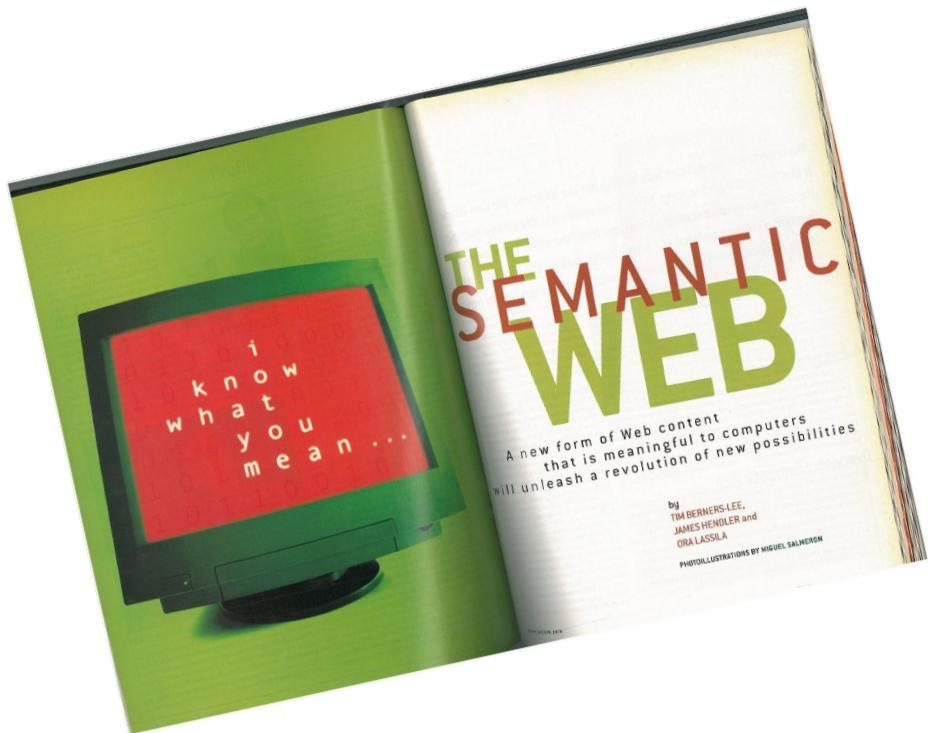
- Digital archives
- Institutional repositories (preprints, postprints, open access publications, ...)
- Digital image libraries
- Research data repositories
- ...



More than 2500 open access repositories worldwide.

Source: The Directory of Open Access Repositories, <http://www.opendoar.org>, retrieved June 06, 2014.

The Semantic Web



„Information varies along many axes. One of these is the difference between information produced primarily for human consumption and that produced mainly for machines. [...]“

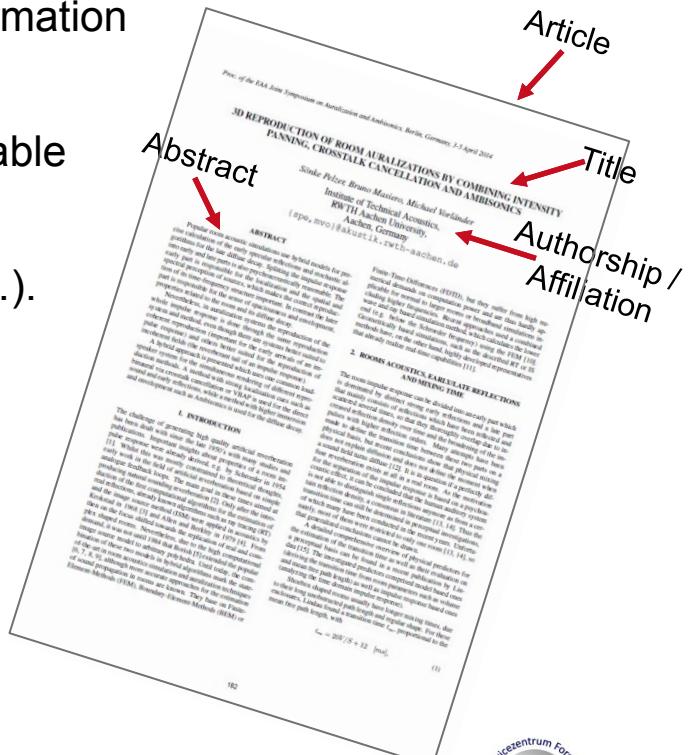
To date, the Web has developed most rapidly as a medium of documents for people rather than for data and information that can be processed automatically.“

Berners-Lee, Handler, Lasilla 2001

Source: Tim Berners-Lee, James Hendler, Ora Lasilla: *The Semantic Web*. In: *Scientific American*, Vol. 284, No. 5, 2001, pp. 28-37.

Semantic Web in a Nutshell

- Information in the web is oriented towards human consumption.
- There is much implicit information: A human understands the context.
- Basic concept of the Semantic Web: Make implicit information explicit, so it can be processed automatically.
- Make information on the Web comparable and combinable even across databases, sites, domains, ...
- Use Web standards (HTTP, HTTPS, SPARQL, RDF, ...).
- Describe domain knowledge in a way it can be used to process information.



The Linked Data Principles

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those things
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so they can discover more things

Tim Berners-Lee

<http://www.w3.org/DesignIssues/LinkedData.html>

Information in Repositories



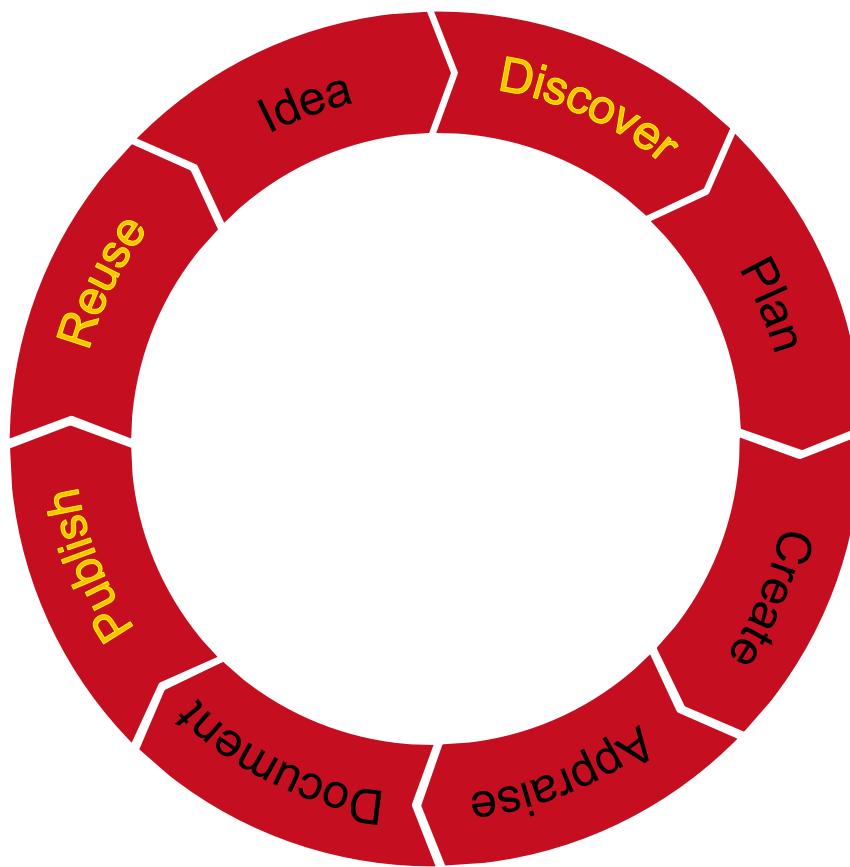
The screenshot shows a detailed view of a publication record:

- Title:** The Impact of the White Noise Gain (WNG) of a Virtual Artificial Head on the Appraisal of Binaural Sound Reproduction
- Authors:** Rasmussen, Eugen; Blau, Matthias; Hansen, Martin; Dodo, Simon; van de Par, Steven; Mellert, Volker; Puschel, Dirk
- Date:** 18-Mar-2014
- Series/Report Nr.:** Proc. of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, 2014/paper #28
- Zusammenfassung:** As an individualized alternative to traditional artificial heads, individual head-related transfer functions (HRTFs) can be synthesized with a microphone array and digital filtering. This strategy is referred to as "virtual artificial head" (VAH). The VAH filter coefficients are calculated by incorporating regularization to account for small errors in the characteristics and/or the position of the microphones. A common way to increase robustness is to impose a so-called white noise gain (WNG) constraint. The higher the WNG, the more robust the HRTF synthesis will be. On the other hand, this comes at the cost of decreasing the synthesis accuracy for the given sample of the HRTF set in question. Thus, a compromise between robustness and accuracy must be found, which furthermore depends on the used sensor (sensor noise, mechanical stability etc.). In this study, different WNG are evaluated perceptually by four expert listeners for two different microphone arrays. The aim of the study is to find microphone array-dependent WNG regions which result in appropriate perceptual performances. It turns out that the perceptually optimal WNG varies with the microphone array, depending on the sensor noise and mechanical stability but also on the individual HRTFs and preferences. These results may be used to optimize VAH regularization strategies with respect to microphone characteristics, in particular self noise and stability.
- Beschreibung:** This contribution was part of the EAA Joint Symposium on Auralization and Ambisonics in Berlin, 2014. Please cite the contribution according to the style: FirstName1 LastName1, FirstName2 LastName2, FirstName3 LastName3: "Title of the Article", in: Proc. of the EAA Joint Symposium on Auralization and Ambisonics in Berlin, 2014. Please cite the contribution according to the style: FirstName1 LastName1, FirstName2 LastName2, FirstName3 LastName3: "Title of the Article", in: Proc. of the EEA Joint Symposium on Auralization and Ambisonics in Berlin, 2014, pp. XXX-XXX, DOI: <http://dx.doi.org/10.14279/depositonce-29>
- URI:** <http://depositonce.tu-berlin.de/handle/11303/186>
<http://dx.doi.org/10.14279/depositonce-29>
- Aparece in:** Proceedings of the EAA Joint Symposium on Auralization and Ambisonics 2014
 Servicezentrum Forschungsdaten und -publikationen

The data stored in repositories are particularly suited to be used in the Semantic Web:

- Metadata already exist in a structured form.
- They do not have to be generated or entered manually for publication as Linked Data.
- “Just” bring the data stored in databases in a form that corresponds to the Linked Data Principles.

Research Data Lifecycle



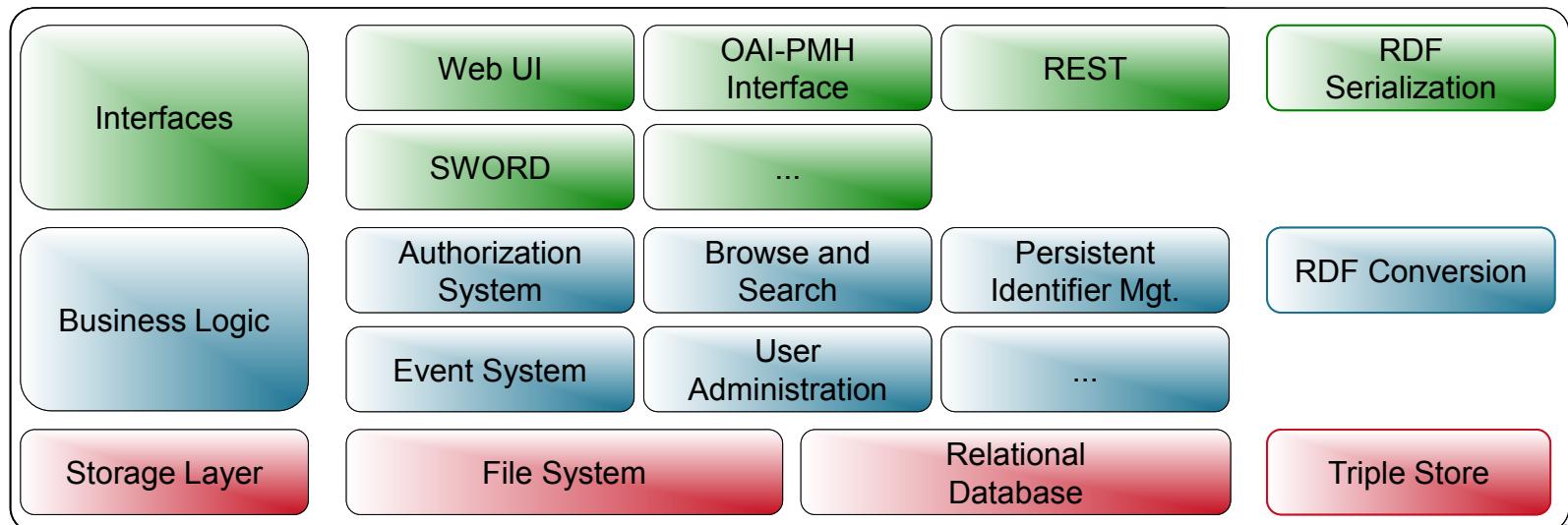
Data exchange with Repositories

- OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting): de facto standard in the context of repositories
- But: limited to that context
- Google retired support for OAI-PMH in 2008 (used before as alternative to the sitemap protocol)
- “Just” an interface, not a format

- Linked Data is a generic, native way of data exchange, not only in the field of repositories (e.g. Wikidata)
- Data published following the Linked Data Principles is self-descriptive
- **Linked Data simplifies data exchange with repositories**

Extending Repositories

- Use Persistent Identifiers in form of HTTP(S) URIs ([http://dx.doi.org/...](http://dx.doi.org/))
- Convert data into RDF and add links!
- Convert public data only
- Store converted data in a Triple Store (RDF database)
- Use the Triple Store as SPARQL endpoint and cache
- Export data in appropriate formats (RDF/XML, Turtle, ...)



DSpace 5

- DSpace is the most often used software for open access repositories worldwide
- Release of DSpace version 5.0 planned for December 2014 (this year!)
- Will contain support for Linked Data (RDF/XML, Turtle, N-Triples, SPARQL)
- Will support content negotiation
- Highly configurable, good default configuration included

**If you're about to use DSpace 5.0 or above
please consider switching Linked Data Support on.**

Technische Universität Berlin
Universitätsbibliothek
Pascal-Nicolas Becker
p.becker@tu-berlin.de
Tel.: +49 30 314-76345

Servicezentrum Forschungsdaten und –publikationen
<http://szf.tu-berlin.de>

Repository DepositOnce
<http://depositonce.tu-berlin.de>

Thesis „Repositorien und das Semantic Web“ (in German)
<http://www.pnjb.de/uni/diplomarbeit/>