ORBITING THE WEB CLOUD

Promises and pitfalls of multilingual vocabularies and vocabulary mappings

CIDOC Conference

Dresden, 9. September 2014

Jutta Lindenthal and Axel Vitzhum

Vocabularies which provide unambiguous, specific, and easy-to-find concepts for indexing and information retrieval are becoming increasingly important in the context of the Semantic Web and Linked Open Data. Since many existing controlled vocabularies were not designed with global networking in mind, some pitfalls have to be anticipated. Omitting the important step of semantic validation can lead to the propagation of errors along relationship chains, as will be shown in some examples. It is concluded that meeting the challenges of multilingual vocabulary work for a global information space requires suitable methodologies, supported by software tools. It will be shown how extensions to an existing vocabulary management platform can supply contextual information from linked data sources which can guide mapping and translation processes, making them less prone to the kind of errors detected in earlier mapping and translation exercises.

digiCULT xTree

Linked Open data
GETTY VOCABULARIES

PARTAGE PLUS

Motivation

Translating and mapping terms or concepts may be difficult and cumbersome. Grasping the notion of a word is even more difficult in information retrieval applications, since the indexing terms are detached from their “natural” syntactic and pragmatic context.

In order to facilitate the task and to avoid rework we propose extensions to a vocabulary management software, xTree (provided by digiCULT-Verbund eG), to semi-automatically assist in the translation or mapping procedure, catching errors in the first place. At this stage the envisaged extensions are mainly intended to be a practical feature which helps to save time, cost and effort.

In the following presentation we will review the steps involved in translation or mapping procedures and show how the xTree-assistant could support these steps.

PARTAGE PLUS

Zrcadlo
©Uměleckoprůmyslové museum, Praha

mirall	ca
zrcadlo	cs
Spiegel	de
mirrors	en
espejo	es
peilit	fi
miroir	fr
zrcalo	hr
tükör	hu
specchi	it
spiegels	nl
speil	no
lustro	pl
espelho	pt
ogledalo	sl
spegel	sv

http://partage.vocnet.org/html/vocItem.php?uri=http://partage.vocnet.org/part00110

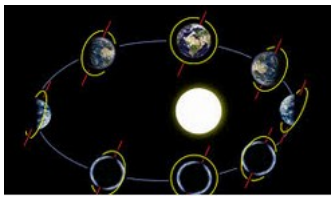
Promises

Linking data and multilinguality are great promises of the Semantic Web, facilitating a common understanding of the knowledge we want to use jointly, promoting interoperability, and fostering sustainability of shared data. Namely, multilingual vocabularies and mappings between authority data are keystones for shared knowledge in the Web. Resources that are provided by different kinds of cultural heritage institutions, and across different countries can be inter-connected and accessed via integrated vocabularies.

The slide above shows an entry of the multilingual vocabulary developed for the Partage Plus project by 25 partners from 17 countries. A sample of Partage Plus objects provided by Europeana is shown here to give an impression how the vocabulary enables a multilingual search in Europeana.

Possible mappings:
AAT: mirrors
GND: Spiegel
LCSH: Mirrors
RAMEAU: Miroirs

knocked off course ...



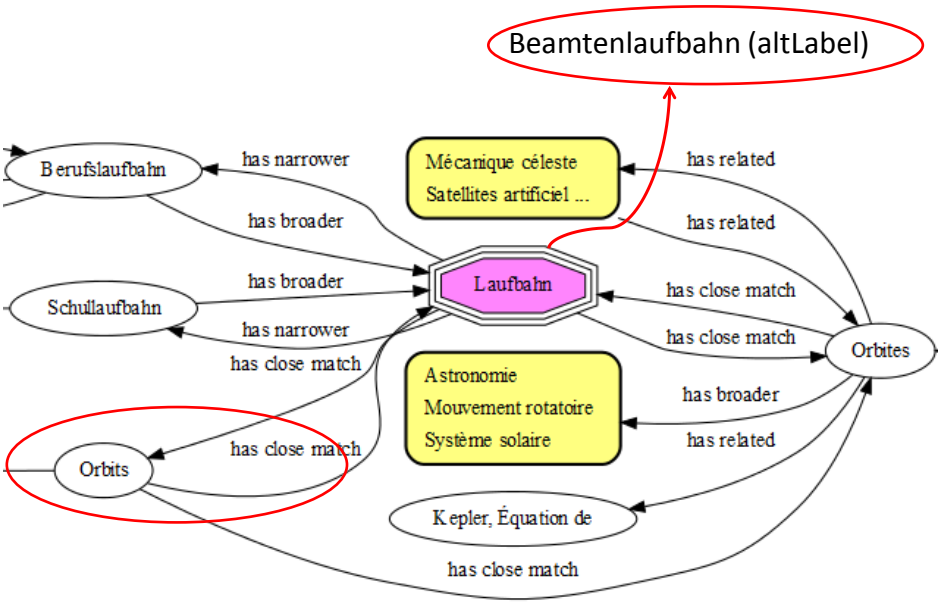
> *“ If you don’t think you have a quality problem with your data, you haven’t looked at it yet. (Jeni Tennison, 2013)*

Pitfalls

However, the world wide linking of (multilingual) data carries a risk: every resource propagates in the Web in a fast and runaway fashion, and any wrong translation or mismatch may end up like a Chinese whisper, misleading the user – or providing for a good laugh.

Imagine a search result like the one shown above. We assume a search with the string *orbit*, and a result showing these three items among others: an image of celestial bodies in an orbit, a running track, and a street named *Beamtenlaufbahn* (civil service career). Only the first item matches the query, the other ones are out of scope. Admittedly, this is a fictitious search result that could nevertheless occur. What is the reason?

Two authority concepts, very different in meaning, have been linked by a close match, and this erroneous link continues to be accessible via the Library of Congress Linked Data Service.



Source: http://semanticweb.cs.vu.nl/europeana/browse/list_resource?r=http://d-nb.info/gnd/4005077-4

Pitfall 1: false mappings

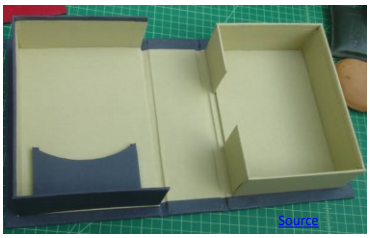
This graphic shows the mapping links between descriptors of the three authority vocabularies Gemeinsame Normdatei (Integrated Authority File, GND), Library of Congress Subject Headings (LCSH) and RAMEAU, as published by the Linked Data Service of the Library of Congress.

The view illustrates, among others, the relations between the LC Subject Heading *Orbits*, and the GND descriptor *Laufbahn* which can refer to either a running track in the sports domain, or the concept *career*. The alternative label *Beamtenlaufbahn* gives a hint to the intended meaning of the descriptor. The mapping erroneously happened probably because of the similarity between the German terms *Umlaufbahn* (orbit), and *Laufbahn* (career).

The visualisation is provided by the AMALGAME project. Amalgame (AMsterdam ALignment GenerAtion MEtatool) is a tool for finding, evaluating and managing vocabulary alignments.

For further information see http://semanticweb.cs.vu.nl/amalgame/

... back on track



Solander box

519. SOLANDER BOX

A box with a hinged top, shaped like a thick book, for storing prints, pamphlets and documents.

DU - overslagdoos (f); solander (m)
FR - boîte à archives (f)
GR - Sammelschachtel (f), Kapsel (f)
IT - scatola a forma di libro (f)
SP - caja en forma de libro (f)
SW - kapsel i form av en bok; dokumentask

Kapsel

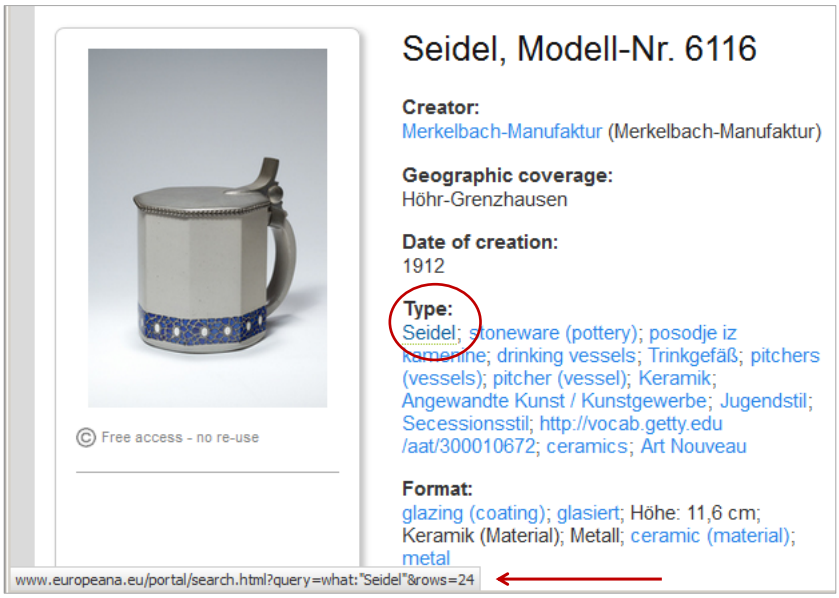
Source: http://archive.ifla.org/VII/s30/pub/mg1.htm#g519

Pitfall 2: false translations

In 1984 the *Section of Art Libraries* of the ***I**nternational **F**ederation of **L**ibrary **A**ssociations and Institutions* (IFLA) developed a *Multilingual Glossary for Art Librarians* in English with indexes in Dutch, French, German, Italian, Spanish and Swedish. The mistranslation of *Solander box* into German *Kapsel* (capsule) migrated to some dictionaries and also slipped into the Art & Architecture Thesaurus. *Kapsel* itself is a homograph, denoting either a pharmaceutical capsule or a space capsule.

The term *Sammelschachtel* does not match the concept *Solander box* either, since it includes all kinds of collectable boxes. The German equivalent for *Solander box* is, one might have guessed, *Solander-Box*.

This example also shows a common pitfall caused by the natural language economy. Often a compound term is shortened to its focus, e.g., *Kapsel* denoting either *Raumkapsel* or *Medikamentenkapsel*.



Source: Europeana Record 2026126

Other pitfalls

This Europeana record illustrates what Linked Data is **not**: the link *Seidel* is a canned text-based query leading to 275 items, including mainly documents referring to persons named *Seidel*, but also to type writers, and last a handful *Bierseidel.*

Another kind of pitfall can be noticed here: the item at hand is not indexed with *tankard*, as similar objects are, but with *pitcher* instead. Thus, this object is dropped from the search result of *tankard,* decreasing recall. This example for an indexing mistake underpins the necessity of clarifying the notion of an indexing term as far as possible, including the usage of example images.

In the following we will have a closer look at the English term *tankard* as part of a controlled vocabulary for information retrieval.

Source: http://babelnet.org/search?word=tankard&lang=EN

Example: what's (in) a tankard?

" The closer the look one takes at a word,
the greater the distance from which it looks back.
Karl Kraus, 1911

Sometimes translating or aligning terms or concepts is straightforward, but more often it turns out to be difficult to find an appropriate equivalent because the meaning of a concept can be ambiguous or may overlap with similar concepts. In order to provide for good retrieval results, the equivalents should be as exact as possible with respect to intension and extension. Inexact interpretations affect the quality of retrieval results: broad equivalents corrupt precision whereas narrow ones worsen recall.

In the following, we will take the English term *tankard* as an example. At first glance its translation may not appear to be a big deal. However, language economy, cultural differences, or regional variants, such as *Krug* and *Seidel*, are challenges for the translation or mapping procedures.

Encircling the concept step by step

Quasi the same by any other name

The meaning of a concept is determined by its context. In information retrieval applications, the context is conveyed by synonyms, broader concepts defining its type and essential properties, narrower concepts giving account of the extension, related concepts delineating semantic boundaries, and definitions describing further characteristics. Checking all these elements manually is cumbersome and time consuming. Hence we propose a computer-assisted procedure.

Step 1: Examine translations back and forth

Step 2: Exclude homographs

Step 3: Check back the source vocabulary

Step 4: Consider semantic relations

Step 5: Ponder the characteristics of the concept

Step 6: Look at real world examples and images in databases

dict.cc
Deutsch-Englisch-Wörterbuch

Wörterbuch Englisch → Deutsch: **tankard** — Übersetzung **1 - 11**

	Englisch	Deutsch	
edit	NOUN a tankard | tankards	–	
	tankard	Krug {m}	65
	tankard	Kanne {f}	7
	tankard	Humpen {m}	5
	tankard	Trinkkrug {m} mit Deckel	

Wörterbuch Englisch ← Deutsch: **krug** — Übersetzung **1 - 16**

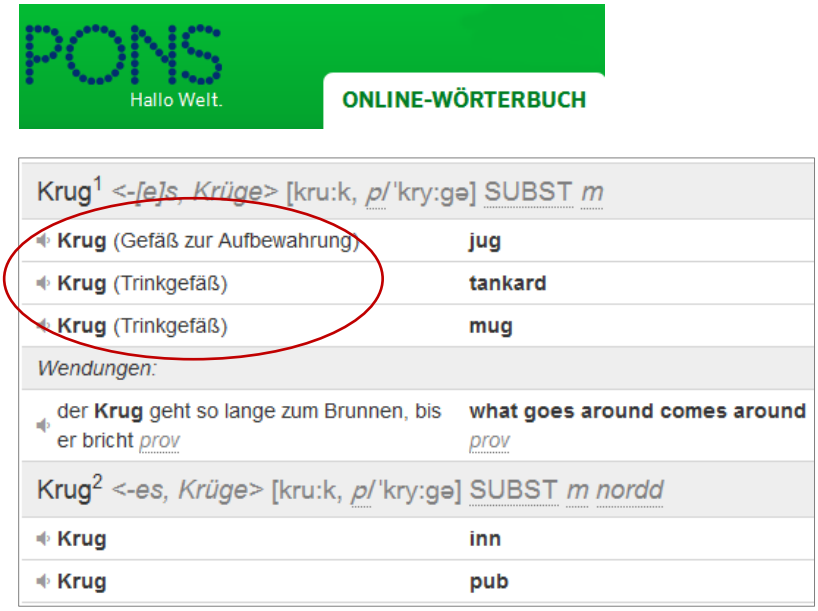
	Englisch		Deutsch	
	–		NOUN der Krug | die Krüge	
	jug		**Krug** {m}	2237
	jar		**Krug** {m}	228
	pitcher [jug]		**Krug** {m}	148
	mug		**Krug** {m}	122
	tankard		**Krug** {m}	61
	pub	gastr.	**Krug** {m} [nordd.] [Schenke, Wirtshaus]	

Source: http://www.dict.cc/

Step 1: Translations back and forth

Most dictionaries, for instance dict.cc, suggest *Krug* as the German equivalent for the English term *tankard* in the first place. However, there are more terms provided for choice. Since it is not granted, that the first suggestion *Krug* is actually the most adequate one, further cues are needed.

A back-translation can give advice, since it often detects further English equivalents. In this example the back-translation reveals that there is a good many of possible equivalents for *jug*, where *tankard* is only a less common option. Instead, the most common translation of *Krug* is *jug,* as the number at the right indicates. The numbers are statistics from a vocabulary trainer and show the frequency of a suggested translation.



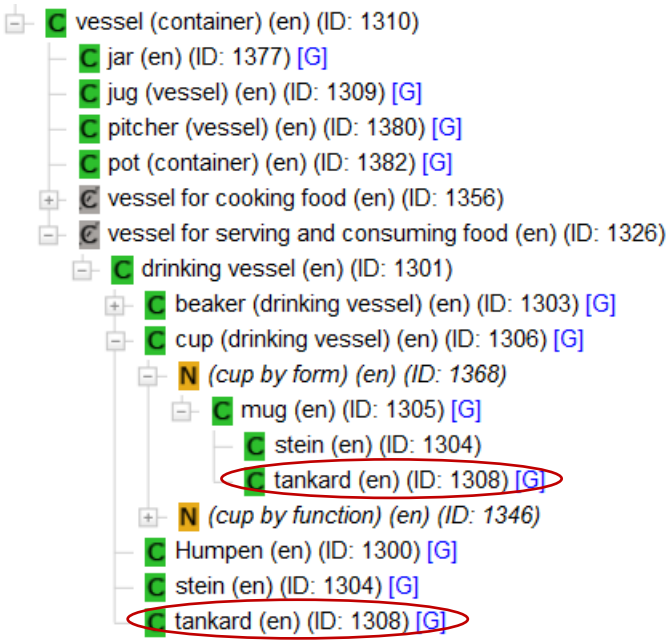
Source: http://de.pons.com/

Step 2: Exclude homographs

In natural language, homographs do occur more often than commonly expected. They are likely to be overlooked, because users often only bear the meaning in mind they are focussed on. Homographs lead to unwanted retrieval results and information overload.

Monolingual dictionaries, such as the German DUDEN or English Merriam-Webster, as well as multilingual dictionaries usually provide different entries for homographs and further distinguish between homonyms or polysems.

The PONS multilingual dictionary, e.g., distinguishes between the two notions of *Krug* by adding terms which clarify the intended meaning. This makes it clear that the polysem *Krug* can refer to either a drinking vessel, such as a *tankard* or a *mug*, or a storage vessel, best interpreted as a *jug*.



Source: snippet from the tree view for *tankard* in xTree

Step 3: Check back the source vocabulary

A look back into the source vocabulary reveals that almost all of the suggested translations for *Krug*, i.e. *jug*, *jar*, *pitcher*, *mug* are entries in their own right. The record set *jug* is best interpreted by the German term *Krug.* In turn, *Krug* is excluded from being the appropriate correspondence for *tankard*.

Such a „tour of terms“ may appear laborious. However, it saves time in the long run, because rework is avoided in the first place. Otherwise, the false translation of *tankard* into *Krug* may not be detected until *jug* has the turn to be translated. Now a whole series of additional steps have to be taken from deleting the wrong translation through to inserting the right ones instead.

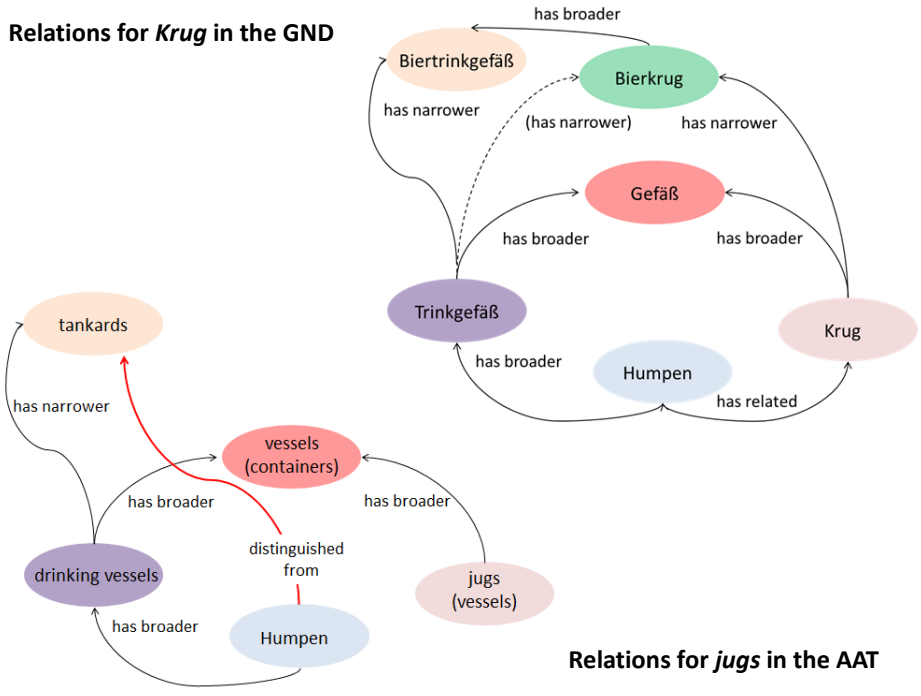
Step 4: Consider semantic relations

The context of semantic relations, especially the hierarchical context, is crucial for at least two reasons:

1. it conveys important information about the kind of an object and its essential properties, and
2. it has critical impact on the results yielded by search expansion.

An accout of immediate parents and children often may suffice to sort out some of the translations in question. E.g., a narrower concept *plates (dishes)* of the concept *dishes (vessels)* excludes a translation into German *Schüssel* as the only equivalent.

However, sometimes it may be necessary to go further up or down in the hierarchy in order to get more hints about the meaning of a concept. E.g., the broader concepts *mug* and *cup* as part of an additional hierarchy in the Art & Architecture Thesaurus further ascertain the characteristic of a *tankard* to serve as a drinking vessel.



Step 4: Consider semantic relations continued

For mapping as well as translation purposes it is particularly illuminating to compare the hierarchical context of a concept in the source and target vocabularies. As a rule of thumb it can be assumed that two concepts are all the more similar in meaning the more equivalent relations they share.

In this example, the AAT and GND concepts *jugs* and *Krug* respectively share the broader concept *vessels* (*Gefäß*). Neither concept does have any hierarchical relationship to *drinking vessels* (*Trinkgefäß*) which in turn is an immediate child of *vessels (Gefäß)* in both of the vocabularies. Only the associative relationships for *Humpen*, a kind of *drinking vessel*, differ in that the AAT links it to *tankard* whereas the GND concept is related to *Krug*. However, these relationships do not contradict each other and thus do not influence the semantic correspondences between the concepts shown here. These correspondences suggest the assumption that *Biertrinkgefäß* or *Bierkrug* are suitable candidates as equivalents for *tankard.*

concept	form	features	material	function/origin
Krug ***Duden***	- cylindrical - bulbous	1-2 handle(s)	- stoneware - glass - porcelain	storing, transporting, serving liquids
jugs (vessels) ***AAT***	- often large capacity - narrow neck - sometimes pouring lip	one handle	- earthenware - stoneware - porcelain	not specified
tankards ***AAT***	- tall	- one handle - usually a lid - often a thumbpiece	- generally silver - or pewter - sometimes glass	drinking
Humpen ***AAT***	- very large - cylindrical - often decorated	- usually a handle - often a lid - often a thumbpiece	- usually glass - ceramic - metal	beaker or stein, usually for drinking beer or wine, Germany from the 16th century onwards
Humpen ***GND***	- cylindrical - bulbous	- lid - handle	various	not specified

Step 5: Ponder the characteristics of the concept

Definitions and scope notes describe important characteristics of concepts. Usually, the description of museum objects regards form, function, material or place and time. These statements should be considered thoroughly, since they often give the decisive clue where to draw the line between two objects. Concepts wich cannot be exactly told apart impede indexing and retrieval quality. The notes below are from the Art & Architecture thesaurus:

tankards

Note: Tall, one-handled drinking vessels, usually with a hinged lid and often a thumbpiece; generally made of silver or pewter, but sometimes made of glass. http://vocab.getty.edu/aat/300043256

jugs (vessels)

Note: Vessels, generally made of earthenware, stoneware or porcelain and often of large capacity, which have a narrow neck and a handle (usually a vertical loop or scroll handle); may sometimes have a pouring lip. http://vocab.getty.edu/aat/300045685

Unknown
Cypriot, 2300 - 1650 B.C.
Terracotta
17 3/8 in.

Jug with a Man and Deer
Source: The J. Paul Getty Museum

Unknown
Possibly northern Bohemian, 1578
Free-blown colorless (brownish-green) glass with enamel and applied decoration
H: 10 5/8 in.

Covered tankard
Source: The J. Paul Getty Museum

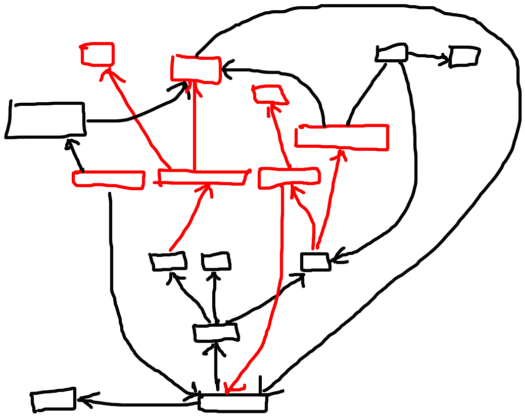
Step 6: Look at real world examples and images in databases

These images of a jug and a tankard, provided by the J. Paul Getty Museum, illustrate some essential differences between the two types of objects. Note that the image *Jug with a Man and Deer* is given as a representative example for *jugs (vessels)* in the Art & Architecture Thesaurus.

A search for images in online portals, e.g., *Deutsche Digitale Bibliothek* or *Wiki Commons*, gives an account of the actual idea of the object in question. Dipping into Google Images can give advice about the popular understanding of a term and sometimes also reveals further homographs.

Recognising the precise notion of a concept is a way that prepares vocabularies for the challenges of the Semantic Web, also with respect to trustworthiness and reliablity. Designing workflows that allow for continuous maintenance and review, using all available tools and instruments sketched out here, will contribute to this end.

Mockup of envisaged extensions of xTree



For information on the envisaged extensions of xTree please see http://www.digicult-verbund.de/files/xtree/CIDOC_20140909.pdf

Conclusion and future work

" Instead of solving interoperability problems we should rather try to prevent them.
Eero Hyvönen (2010)

- survey available vocabulary resources for reliabiliaty and programmatic access (API)
- evaluate existing alignment tools, such as AMALGAME, and matching algorithms (e.g., by Navigli et al.) for re-use
- survey evalation tools and quality checkers (e.g., by OAEI)
- evaluate if Formal Concept Analysis could be sensibly employed
- evaluate where visualisation of concept graphs is practically useful for exploring semantic context

References

- Hyvönen, Eero (2010):
  Preventing Ontology Interoperability Problems Instead of Solving Them.
  http://www.semantic-web-journal.net/sites/default/files/swj33.pdf
- Pilehvar, Mohammad Taher; Navigli, Roberto (2014):
  A Robust Approach to Aligning Heterogeneous Lexical Resources.
  http://www.pilevar.com/taher/pubs/ACL_2014_Pilehvar_Navigli.pdf
- Tennison, Jeni (2013): Five Stages of Data Grief.
  http://theodi.org/blog/five-stages-of-data-grief
- Dictionaries and Vocabularies
  - Art & Architecture Thesaurus http://www.getty.edu/research/tools/vocabularies/aat/
  - BableNet – http://babelnet.org/
  - dict.cc – Online-Wörterbuch Englisch-Deutsch http://www.dict.cc/
  - Duden online – Wörterbuch http://www.duden.de/woerterbuch
  - Multilingual Glossary for Art Librarians http://archive.ifla.org/VII/s30/pub/mg1.htm
  - PONS – Online-Wörterbuch http://de.pons.com/
  - WikiTerm – Terminology Research and Extraction Tool http://wikiterm.uebersetzer.org/
    Part of MetaTerm – Online Terminology Search http://www.uebersetzer.org/

" Given enough eyeballs, all bugs are shallow.
Linus's Law by Eric Raymond, 1999

For further information please contact

Jutta Lindenthal, Independent Information Consultant, Lübeck.
Phone: 0049 (0)4502-8809421.
E-Mail: jutta.lindenthal@gmail.com

Axel Vitzthum, Head of IT development, digiCULT-Verbund eG Kiel.
Phone: 0049 (0)431-908914-73.
E-Mail: axel.vitzthum@digicult-verbund.de