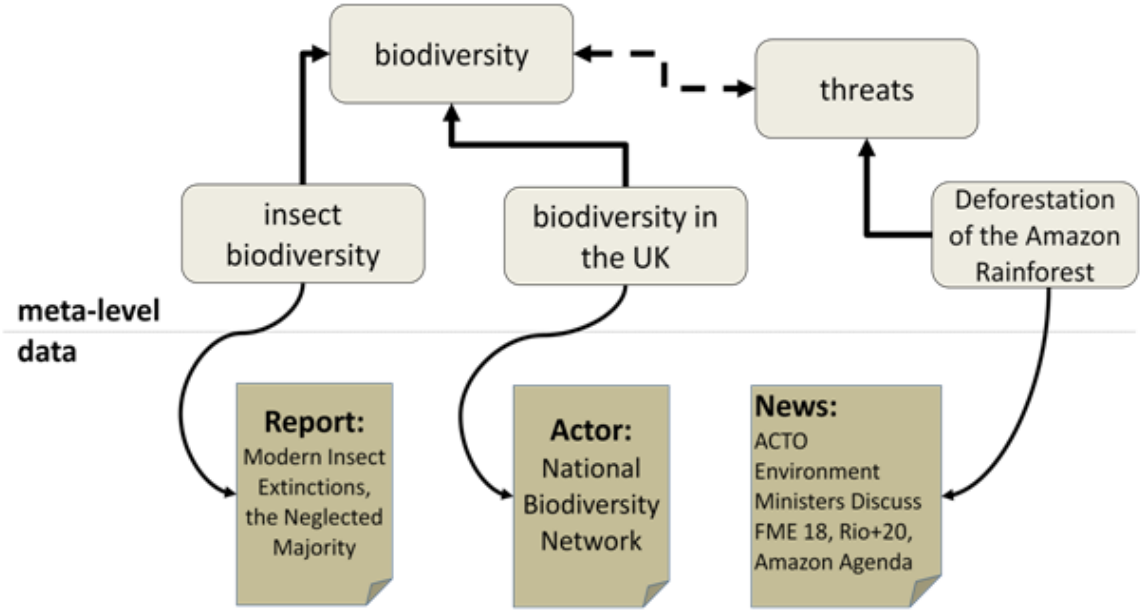
5 arguments in favour of thesauri & controlled vocabularies

Make knowledge explicit and available throughout your department, your organization, your partner network, the whole world!

Knowledge is quite often well hidden in our heads. Every day we have to ask around, write emails, search in documents etc. to put together an information jigsaw puzzle using the information from different websites, internal sources, mails etc.
Creating a thesaurus and establishing this knowledge network as a valuable source for ongoing research activitites will help to make existing knowledge re-usable. Thesauri represent associations between concepts and their names. When using a thesaurus to acquire new knowledge one can "browse" on the meta-level first and can systematically dig deeper and deeper into a domain by exploring data and documents. Thesauri are very much alike the way human beings think. We don´t have tables or primary keys in our minds, but rather abstractions, categories, specialisations, is-part-relationships and different names for the same thing, in our own language and for all the other languages we can speak or understand.



Dividing knowledge into meta-level knowledge and data-level knowledge is similar to the way we organize our individual knowledge about a certain domain: We don´t 'store' all the attributes of our neighbour´s dog in our mind, instead we rather try to explain properties of certain instances of dogs by their category, e.g. 'Rusty is a Scottish Terrier. This makes him a good watchdog'. Since we know that Scotties are territorial, alert, quick moving and feisty in general, we don´t have to remember all these things for all of the Scotties in our neighbourhood, we just have to remember that Rusty and Boris are Scottish Terriers. So we will always be careful when we enter our neighbour´s garden although "Rusty" has always been friendly to us. Nevertheless, by seperating meta-level from data-level our knowledge can be spread out more efficiently. It´s interesting for many of us how Scottish Terriers behave in general but it´s not interesting for most of us that "Rusty" is afraid of red balls.

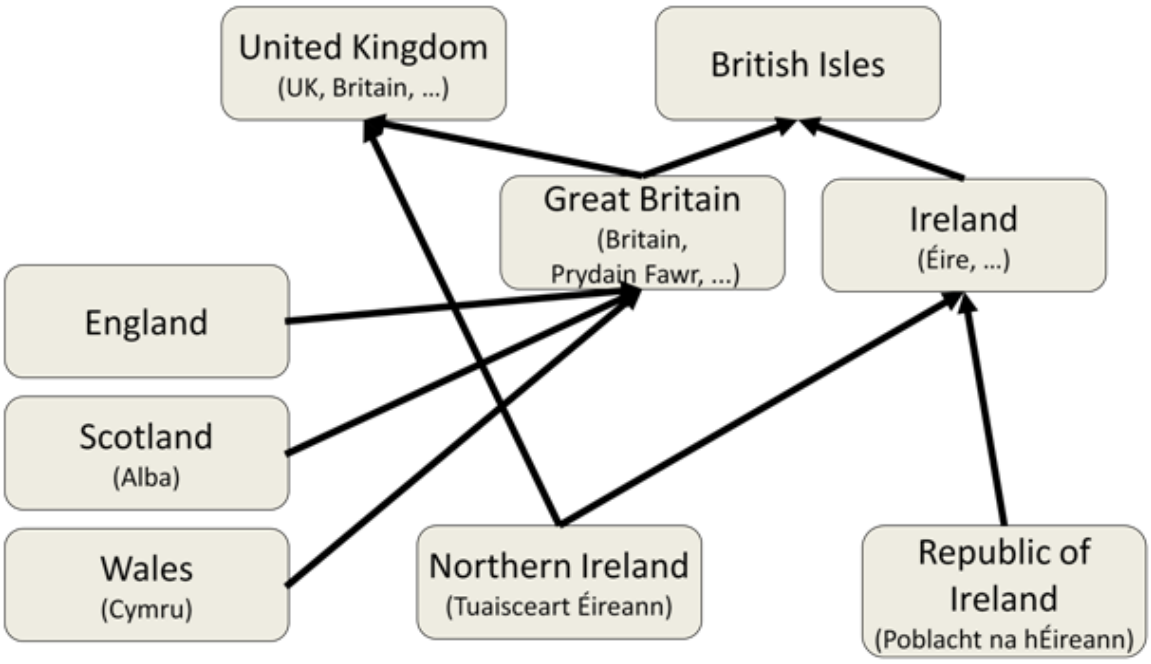
Conquer the Babylonian language confusion!

Talking about complex systems and contexts is always about defining words and their meanings. Even team members who have worked together for years and years are not always on the same page because of changes in terminology and multiple translations for various word meanings. This is even more relevant in organisations working in multi-lingual environments on a global scale.

Thesaurus management is not necessarily about defining a 'gold standard' in a sense of stipulating which words have to be used to express a certain idea. It´s rather about collecting different ways to express a certain meaning, it´s about structuring the relations between words and the concepts behind them and the relationship between those concepts.

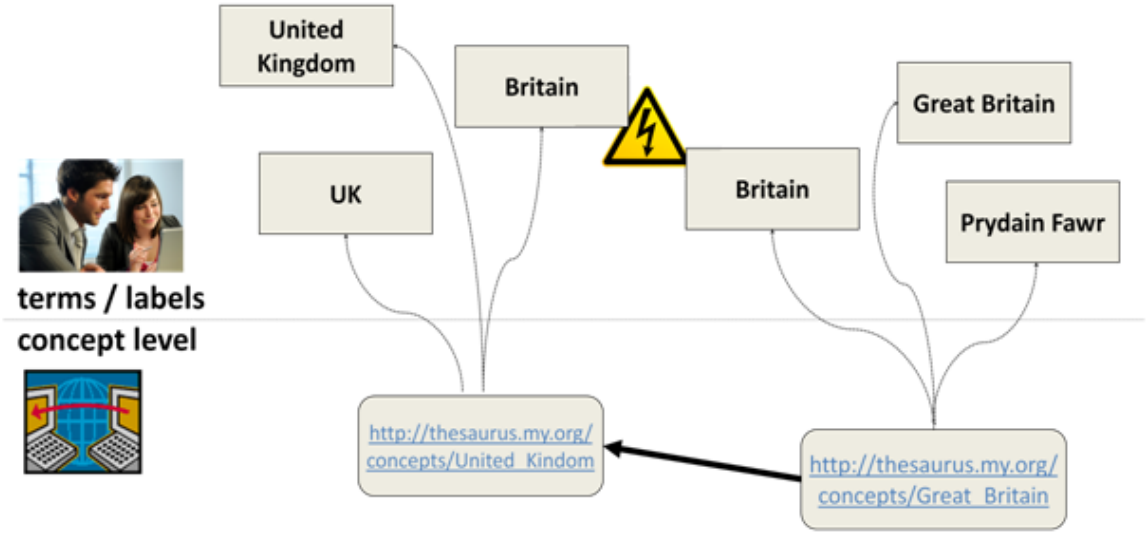
The provision of thesauri to a handy size, for instance on a website or on the intranet (e.g. http://thesaurus.my.org/) can support people (e.g. new members and employees) to understand better what all the abbreviations and specialist terms mean. Conquering the Babylonian language confusion also means to fight the substantial amount of half-knowledge which usually is all around.

To illustrate a common variety of language difficulty, let´s take a look at the following question: "Is Irish an official language in Britain?"



As we can see "Britain" can be used to refer to "Great Britain", the isle, or it can also be used to refer to "United Kingdom", the sovereign state. If we mean "UK" by saying "Britain" the answer to the question above is "yes", if we mean "Great Britain" the answer is "no".

To solve this problem (we call it "disambiguity" in linguistics), concept based thesauri make a difference between the terms we use and the concepts 'behind' the terms. This method allows to build interfaces which help users to refine their search query. If someone searches for "Britain" the system will ask what exactly is meant by "Britain" and will offer two options: the isle and the sovereign state. Concepts on the Semantic Web can be adressed via their "Uniform Resource Identifier (URI)" and they are connected to all the terms (labels) which are most often used to refer to this concept. As an example, a URI for United Kingdom could be http://thesaurus.my.org/concepts/United_Kindom, clearly differentiable from http://thesaurus.my.org/concepts/Great_Britain referring to the isle.



Help to make search and research activities more efficient!

The above mentioned fuzziness of human language causes serious problems in our daily work dealing with computer systems like search engines etc. When someone tells you that he will go to Java this year, we are instantly able to understand that "Java" refers to an island and not to a programming language, computers don´t. The above mentioned approach introducing URIs to adress concepts and their relations helps to escape from this awkward situation. Autocomplete is a popular search assistant which helps to disambiguate search terms at the very beginning of a search process.

Autocomplete based on a thesaurus helps users to remember all the names of a concept and to pick the proper one.

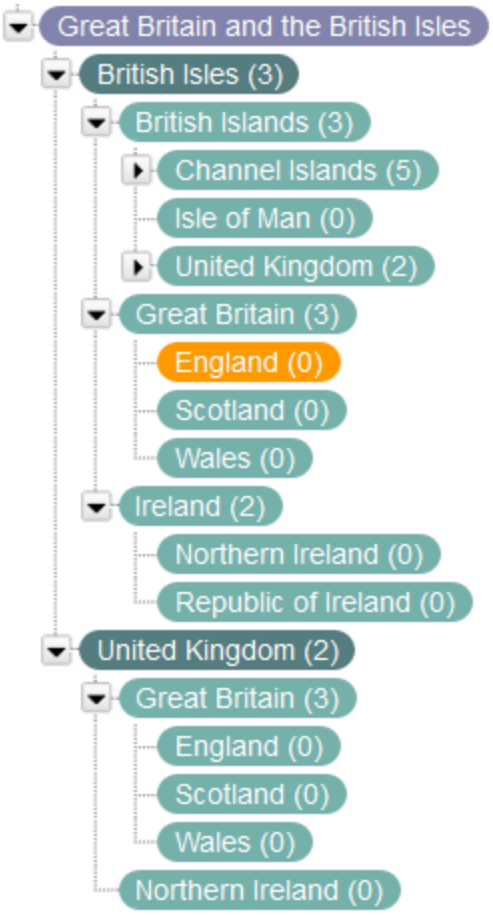
In the background a so called 'query expansion' helps to increase the recall of a search procedure whilst reducing efforts to formulate the proper query: 'mann' will be expanded by 'mann' OR 'isle of man'.

Thesauri as a knowledge network in the back can also serve as search assistants: Search refinements can be offered like "You were looking for 'Isle of Man', are you interested in other 'Britsh Islands' like the 'Channel Islands' or 'United Kingdom'? A comparable example in the field of clean energy can be found at http://www.reegle.info/clean-energy-search.

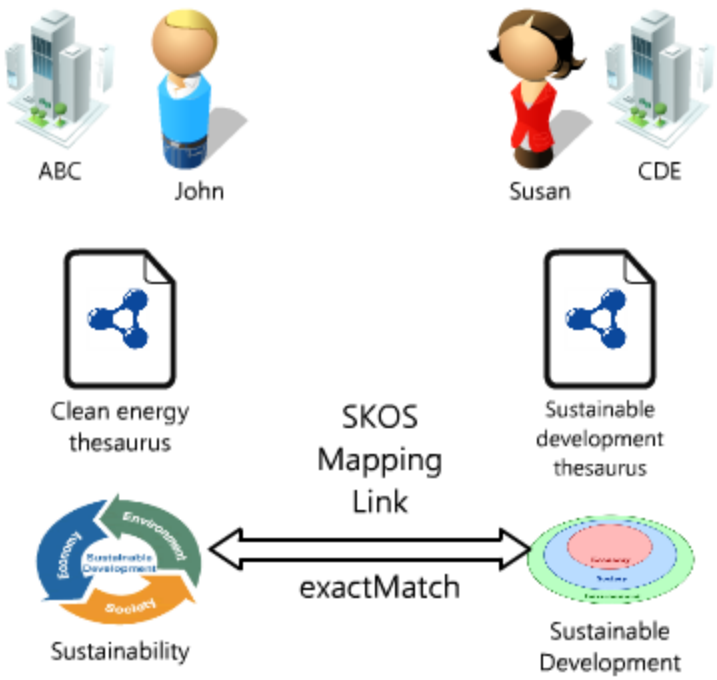
Link all your information sources in a meaningful and standardized way!

The idea of describing knowledge and models of knowledge is nearly as old as humanity itself. Any cave painting describing efficient ways of hunting are means to represent knowledge. Any categorization system used by librarians is built upon methods which were developed throughout history.

The World Wide Web as a place to organize and distribute knowledge is rather young compared to other culture techniques. W3C has published a large set of standards to structure and link information in decentralised environments called the "Semantic Web". A subset of the Semantic Web is called the "Simple Knowledge Organization System (SKOS)" which is meant to be used for thesaurus management. With SKOS thesauri can be generated, published and reused in a standardized manner. The core of the SKOS standard (http://www.w3.org/2004/02/skos/) are so called "Concept Schemes" which contain "Concepts". As already discussed, concepts are represented by URIs and can be linked to each other by means of predefined relation types. The most important relation types are "related", "broader", "narrower" and "exact match". With SKOS we can state in a machine-processable way: http://thesaurus.my.org/concepts/Great_Britain skos:narrower http://thesaurus.my.org/concepts/England, which means that "England" is narrower than "Great Britain".

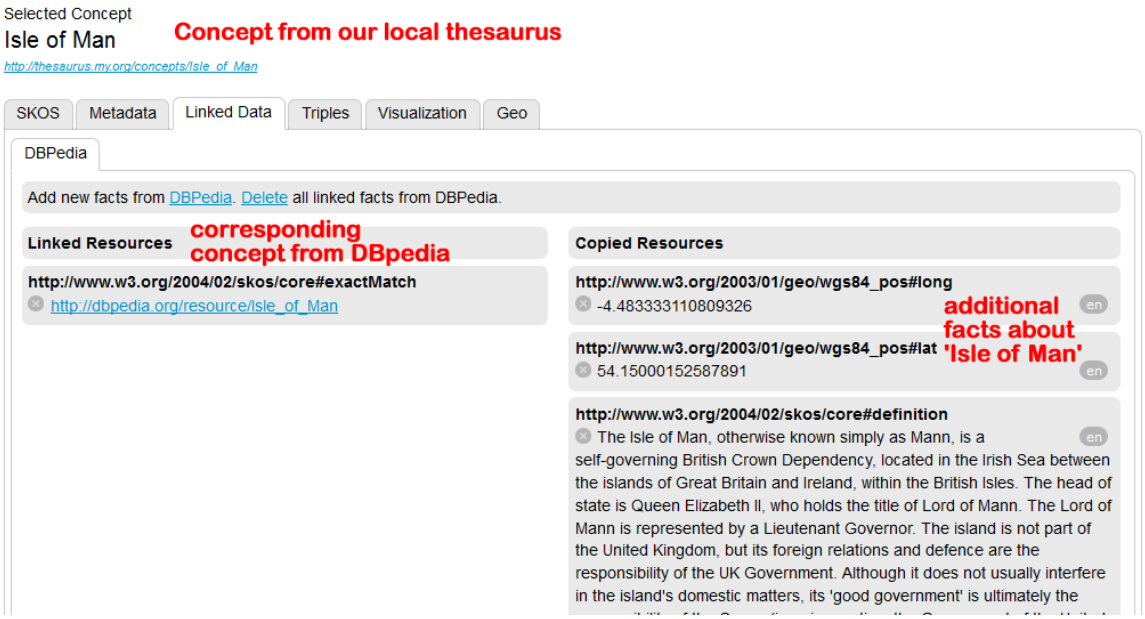


The Semantic Web is the technological basis for a 'web of linked data' (see: http://issuu.com/andreas_blumauer/docs/linked-open-data-essentials). With linked data mechanisms thesauri can be linked to each other even across organisational boundaries. There is no need to develop one single and large thesaurus for a certain domain, it´s rather collaboration that matters in the age of the internet. Imagine, organisation ABC has built a thesaurus in the area of 'clean energy'. It is well established, high-quality and frequently used by a lot of applications. Organisation CDE is specialised in a domain which is related to ABC´s knowledge domain: Sustainable development. CDE wants to start with its own thesaurus and wants to refer to ABC´s existing knowledge model. On top of linked data technologies and SKOS the following knowledge model can be generated:



Using SKOS mappings like skos:exactMatch leaves all concepts at their original place. With mappings different views on the same concept can be integrated into one view and we can even re-use labels and vocabularies of other experts to find their documents without the need to learn their 'language'.

And the very best of linked data is that there is plenty of it freely available on the internet. For example, DBpedia (http://dbpedia.org/) ist the 'semantic sister' of Wikipedia. Nearly everything which you can find on Wikipedia, you can also find on DBpedia, but in a structured, machine processable way. This brings us into the position to 'enrich' all concepts from our thesaurus by additional facts. The following example shows that after having aligned our 'Isle of Man' concept with the corresponding concept from DBpedia additional information like definition, geographical coordinates, images etc. are only one mouse-click away. A very rich knowledge base for a huge variety of domains can be generated by using linked (open) data!



Make skills, interests and knowledge of experts visible!

To search for information means quite often to search for the right person who is experienced in what I´m looking for. Flicking through our adress books and browsing social networks sometimes helps, at least if we start asking around if someone might know someone who ... It is quite obvious that this method is not the most efficient one to identify contact persons for a certain problem. Thesauri as models to describe skills, interests and fields of expertises of people might help to establish a knowledge base which can be used for experts search. Many large organisations use systems called 'yellow pages' to enhance people search but most often such systems lack of proper actuality and topicality. A solution for this dilemma is the introduction of text analysis and social tagging based on thesauri. Whenever we submit reports, minutes or other documents to a system, colleagues will assume that the authors are experts about the topics they are talking about. At least statistically spoken this is correct, so we can analyse all documents written by a certain person and can extrapolate his/her skills and expertise. An example for such an expert search engine based on Twitter can be found here: http://topsy.com/s/sustainable+development/expert

About the author

Andreas Blumauer is CEO of Semantic Web Company (SWC) based in Vienna/Austria. SWC is one of the largest providers of linked data and semantic web technologies in Europe and serves customers like Roche, Credit Suisse, Wolters Kluwer, Texas State University or Biogen Idec. SWC is also one of the pioneers in implementing linked data technologies in enterprises. Andreas is responsible for the strategic development of the PoolParty product family at SWC. In the course of many industry projects he has gained long-time experience in the areas of social media in enterprises, knowledge modelling, text mining, data integration and search based applications.