ZBW
Leibniz-Informationszentrum
Wirtschaft
Leibniz Information Centre
for Economics

Controlled vocabularies in a repository environment: Overview and state-of-the-art

COAR
Confederation of Open Access Repositories

Dr. Timo Borst
Dept. Information Systems and Publishing Technologies
ZBW - Leibniz Information Center for Economics

COAR Webinar, 4th October 2016

Contents

- Definition and scope of CV
- Management, maintenance & governance
- CVs and their integration into repository systems
- Case study: ZBW‘s Standard Thesaurus for Economics (STW)

Definition and scope of CV

Controlled vocabulary are selected lists of terms and phrases (with guidelines for their use), that are used to populate metadata elements. The over-riding goal in using controlled vocabulary is to make the retrieval of resources and information through searches more efficient. Controlled vocabulary reduce ambiguity in language and help to ensure data consistency.

https://www.jisc.ac.uk/guides/metadata/controlled-vocabulary

Definition and scope of CV

Controlled Vocabularies are one of the semantic glues that binds open access repositories and scholarly communication infrastructures together. They offer great benefit to the community, because they ensure interoperability between repositories and repository content, and facilitate greater discovery, tracking and re-use of research materials.

http://dcevents.dublincore.org/IntConf/index/pages/view/workshop

Definition and scope of CV

A controlled vocabulary is an organized arrangement of words and phrases used to index content and/or to retrieve content through browsing or searching. It typically includes preferred and variant terms and has a defined scope or describes a specific domain.

http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/what.pdf

Definition and scope of CV

A controlled vocabulary is a restricted list of words or terms used for labeling, indexing or categorizing. It is controlled because only terms from the list may be used for the subject area covered by the controlled vocabulary. It is also controlled because, if it used by more than one person, there is control over who adds terms to the list, when, and how to the list. The list could grow, but only under defined policies. Most controlled vocabularies also have some form of cross-references pointing from one or more "non-preferred" terms to the designated "preferred" term. Only if a controlled vocabulary is very small and easily browsed, such as on a single page, might such synonyms be excluded.

http://www.hedden-information.com/taxonomies.htm

Definition and scope of CV

Controlled vocabulary

- Restricted list of words or terms used for labeling, indexing or categorizing any kind of information object
- As a standard rule, they contain references to preferred terms defined by some body corporated

Thesaurus

- Always consisting of a term and its context, the latter defined by broader/narrower terms, related terms and synonyms
- In particular, ‚related terms‘ can be used for mapping to CVs from other domains (without taking concern of their hierarchical systematic)

Definition and scope of CV

Taxonomy

- Representing one big hierarchy (or tree) of (controlled) terms
- In contrast to thesaurus with its broader/narrower/related terms, any taxonomy’s term is hierarchically located

Ontology

- Set of concepts defined with attributes and relationships (to other concepts)
- Strong formal requirements on the consistency of CVs (which can be checked e.g. by means of reasoner)

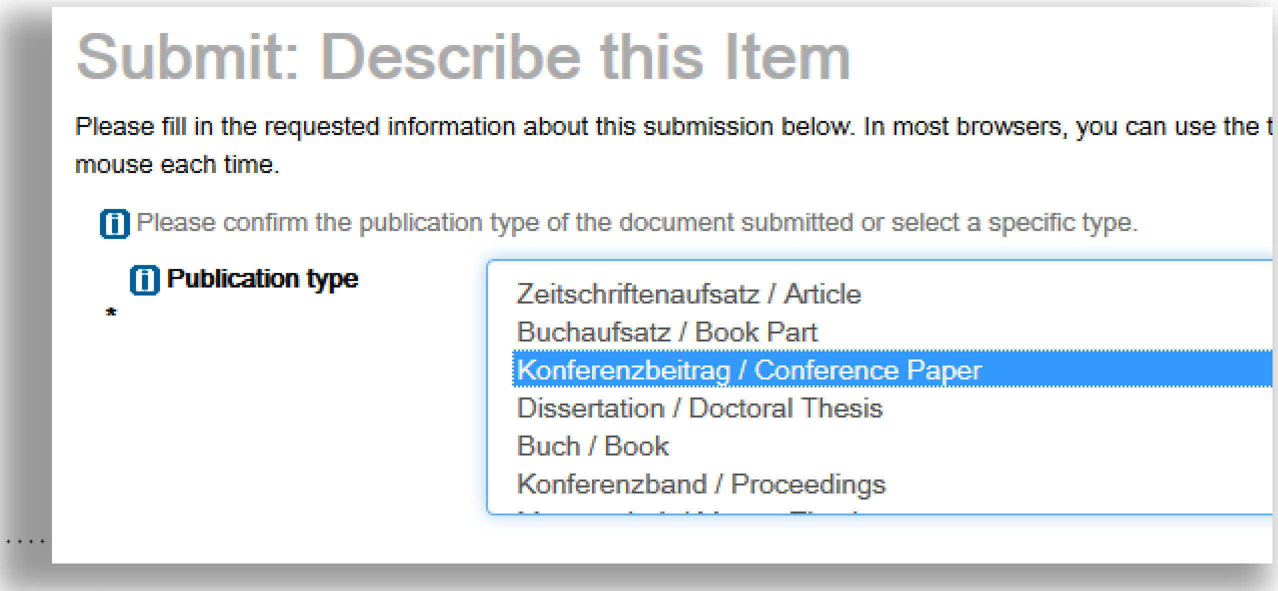
Definition and scope of CV

CVs do not only apply to concepts (which may be the traditional use case), but to any kind of information or research object – except from its title and abstract…

- Author / contributor: controlled resp. preferred names
- Date (e.g., ‚YYYY' according to ISO 8601)
- Document type (e.g., ‚workingPaper' or ‚conferencePaper' according to https://wiki.surfnet.nl/display/standards/info-eu-repo)
- Access rights (e.g. ‚openAccess' in contrast to ‚closedAccess')
- …

Management, maintenance & governance

- Management normally conducted by means of software (either a code fragment within a larger software, or a system particularly designed for the management and maintenance of thesauri)
- Easiest approach: List of predefined metadata values (to support normalization at least at indexing)
- More sophisticated approaches: list of (search/index) terms generated by the indexing/retrieval system



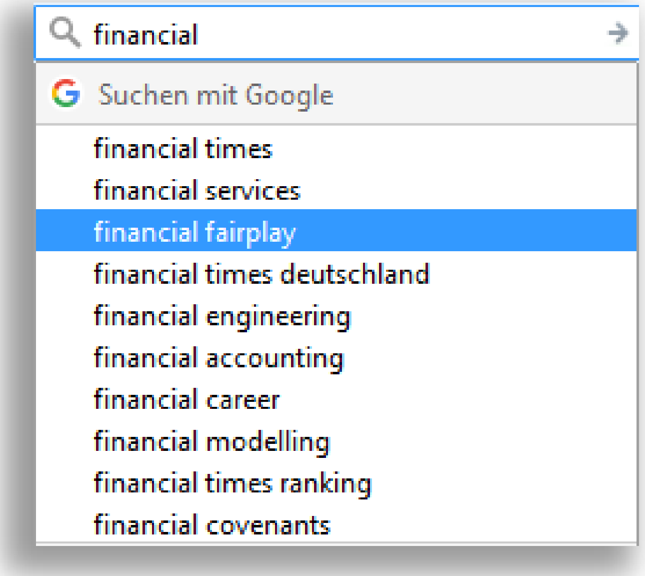
Submit: Describe this Item

Please fill in the requested information about this submission below. In most browsers, you can use the t mouse each time.

Please confirm the publication type of the document submitted or select a specific type.

Publication type *

Zeitschriftenaufsatz / Article
Buchaufsatz / Book Part
Konferenzbeitrag / Conference Paper
Dissertation / Doctoral Thesis
Buch / Book
Konferenzband / Proceedings



financial

Management, maintenance & governance

- Maintenance normally conducted semi-automatically: in the case of subject terms, user input is taken at least as a candidate for enhancing a CV
- Input from either information specialists, or from the regular user (‚crowd‘)
- Workflow at ZBW for maintaining the ‚Standard Thesaurus for Economics‘:
  - New candidates for indexing terms are suggested and introduced as ‚free keywords‘ (and tracked by the thesaurus management system)
  - If a certain amount of documents matches the keyword, it term is integrated into the STW thesaurus becoming either a new descriptor, or a related/broader/narrower term

Management, maintenance & governance

- Governance = a corporate body and/or a set of rules defining the elements of a CV (the latter often implemented by means of software)
- ‚Folksonomy' = explizt absence of governance
- Example of a more sophisticated governance: maintenance of the semantic description of a repository's information objects according to the COAR vocabularies (https://www.coar-repositories.org/activities/repository-interoperability/ig-controlled-vocabularies-for-repository-assets/coar-vocabularies/)
- Workflows needed to introduce, to maintain, to disseminate, to update, to archive and to track CVs

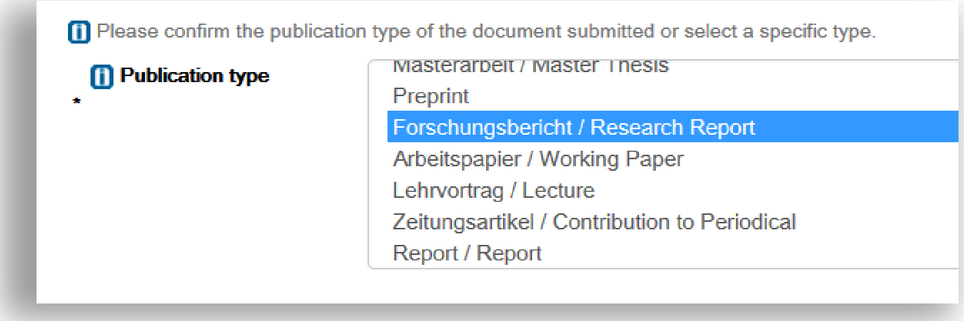
CVs and their integration into repository systems

- In general, Open Source repository packages are reluctant to ship CVs with their distributions (because of legal issues, another issue might be maintenance…)

- Attempts have been made to introduce CVs and Authority Control into repository systems, but in fact there still exist only two approaches…

CVs and their integration into repository systems

1. Dropdown list of predefined values shipped with the software or added to the local deployment

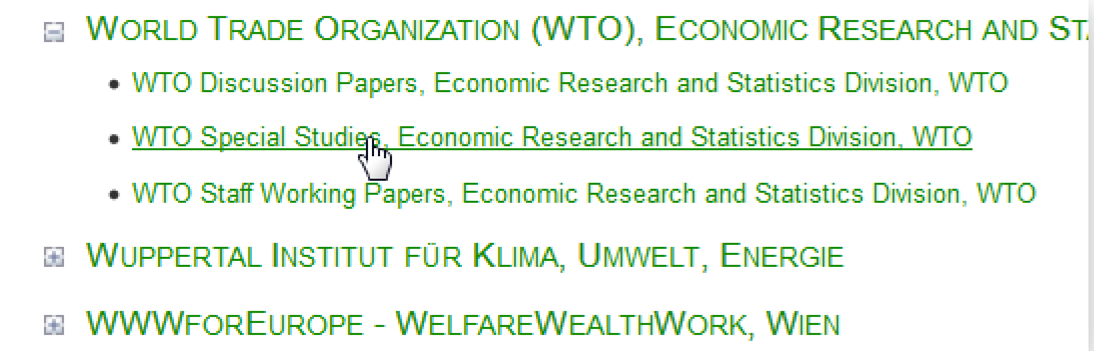


Please confirm the publication type of the document submitted or select a specific type.

Publication type *

- Masterarbeit / Master Thesis
- Preprint
- Forschungsbericht / Research Report
- Arbeitspapier / Working Paper
- Lehrvortrag / Lecture
- Zeitungsartikel / Contribution to Periodical
- Report / Report

2. (XML-)File to be loaded into the repository system, from which a hierarchical treeview will be automatically generated



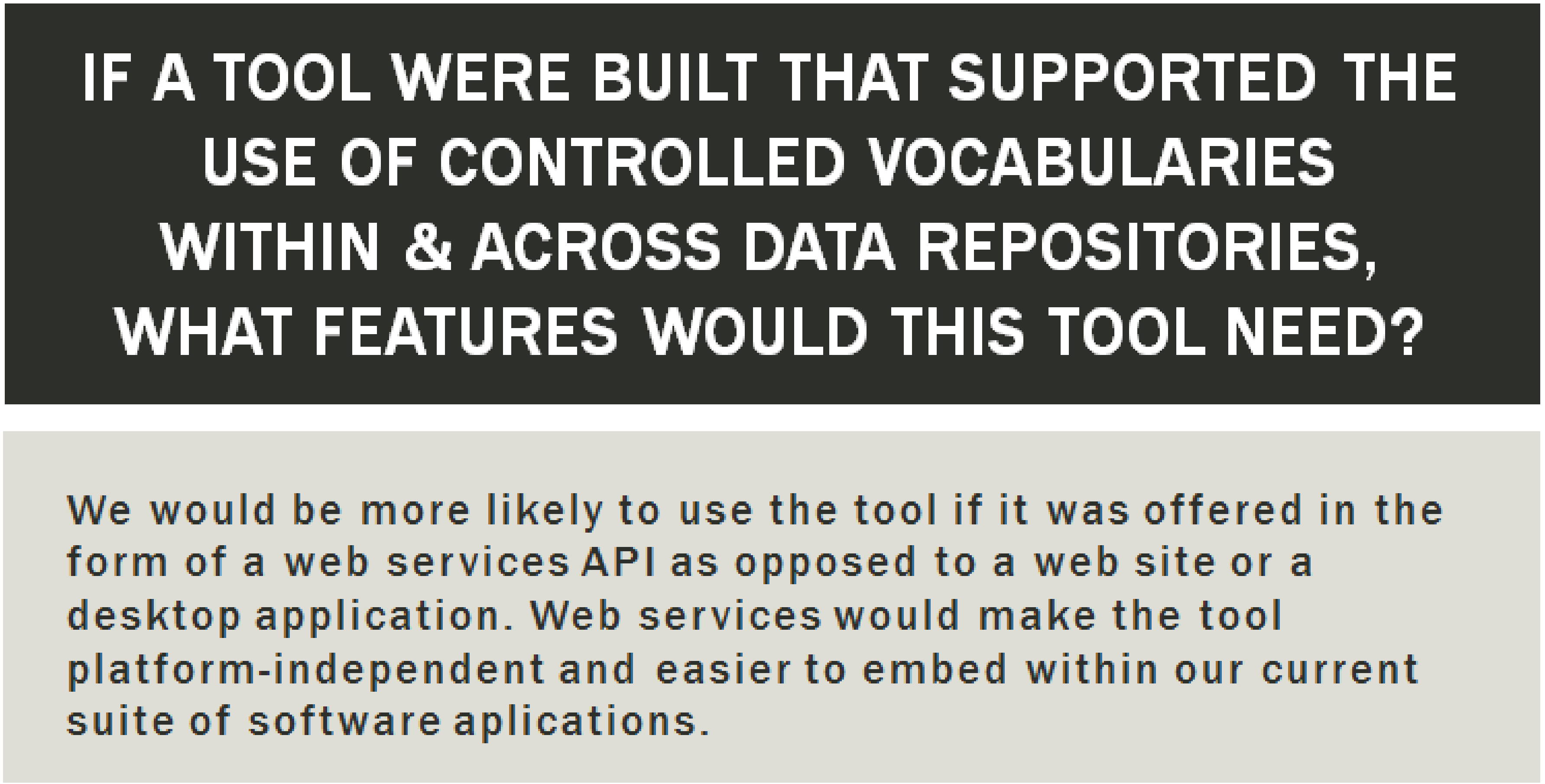
⊟ World Trade Organization (WTO), Economic Research and St

- WTO Discussion Papers, Economic Research and Statistics Division, WTO
- WTO Special Studies, Economic Research and Statistics Division, WTO
- WTO Staff Working Papers, Economic Research and Statistics Division, WTO

⊞ Wuppertal Institut für Klima, Umwelt, Energie

⊞ WWWforEurope - WelfareWealthWork, Wien

CVs and their current integration into repository systems



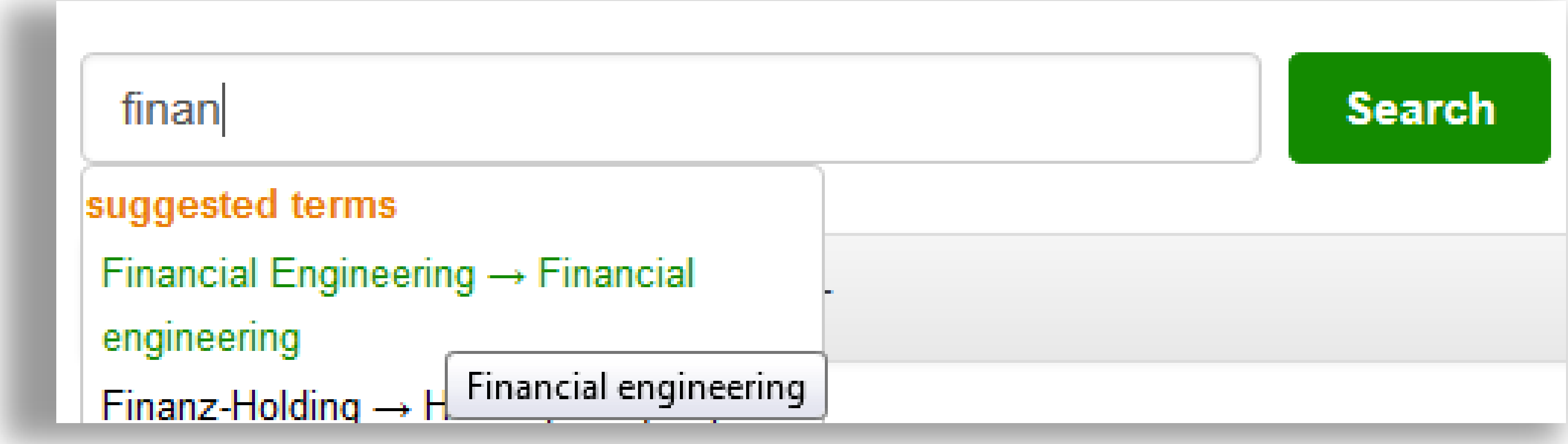
IF A TOOL WERE BUILT THAT SUPPORTED THE USE OF CONTROLLED VOCABULARIES WITHIN & ACROSS DATA REPOSITORIES, WHAT FEATURES WOULD THIS TOOL NEED?

We would be more likely to use the tool if it was offered in the form of a web services API as opposed to a web site or a desktop application. Web services would make the tool platform-independent and easier to embed within our current suite of software aplications.

C. Rowell / J. Greenberg: Controlled vocabulary status & potential in data repositories. Authority Control Interest Group ALA Annual 2013 http://connect.ala.org/node/209259

CVs and their integration into repository systems

- (External) Webservice to be queried and integrated into a repository‘s workflows (e.g., for suggesting subject terms)
- Lightweight approach: CV has to be maintained elsewhere, only access via API must be provided
- Webservice scalable, up-to-date and technically independent from any particular repository platform


finan

suggested terms

Financial Engineering → Financial engineering

Finanz-Holding → H

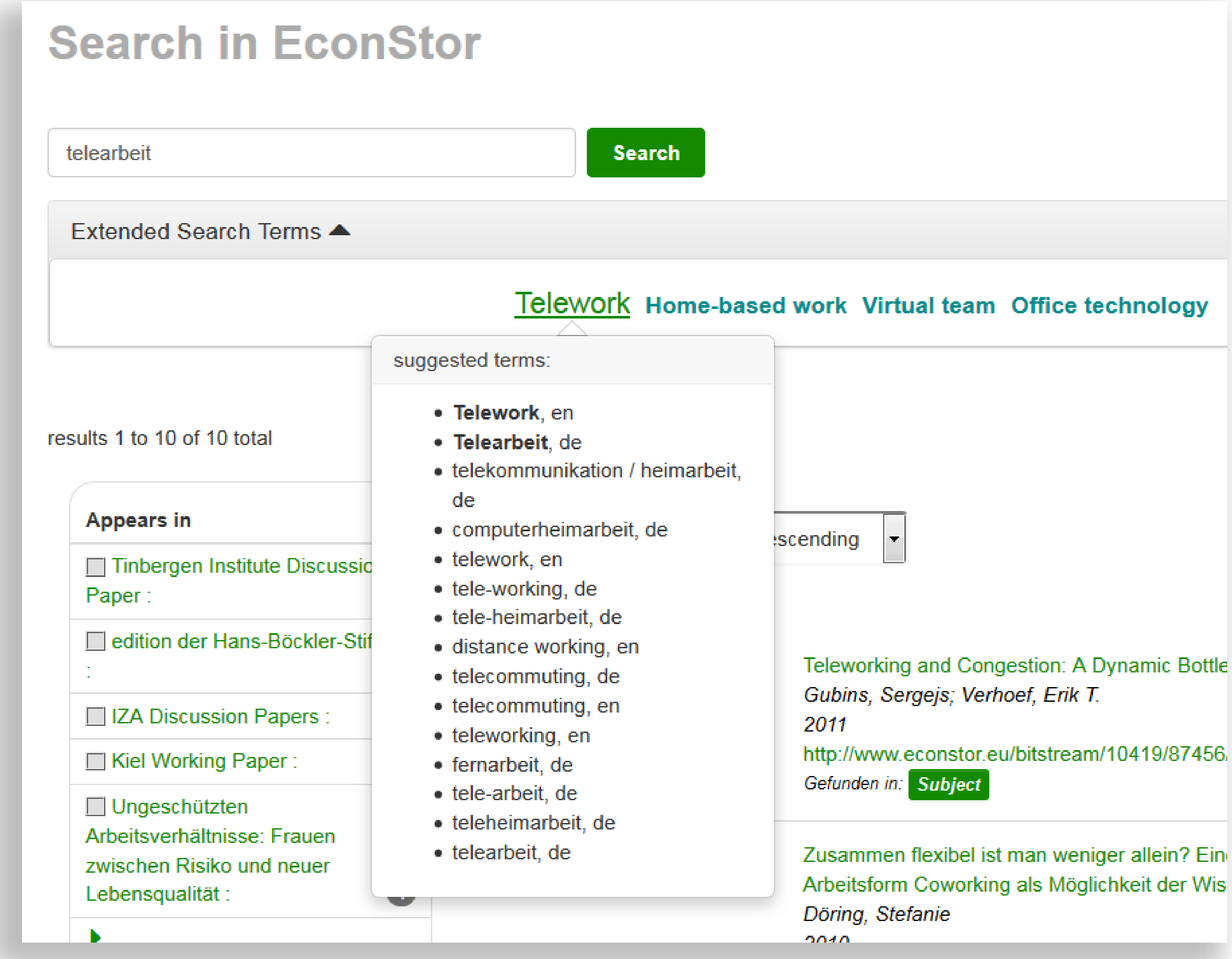
Financial engineering

Case study: ZBW‘s Standard Thesaurus for Economics (STW)

- Created in the 1990s, currently maintained and enhanced by ZBW
- More than 6,000 descriptors in English and in German
- More than 16,000 broader/narrower relations
- 11,000 related concepts
- Mappings to other CVs (GND, DBPedia, Agrovoc,
- Constantly updated and versioning (currently 9.029)
- Available as
  - Website http://zbw.eu/stw/version/latest/about.en.html
  - Downloadable SKOS-Dataset (http://zbw.eu/stw/version/latest/download/about.en.html)
  - Webservice (http://zbw.eu/beta/econ-ws/about)

Case study: ZBW‘s Standard Thesaurus for Economics (STW)

- Webservice to support both indexing and searching
- Autosuggest resp. Autocompletion
- Term expansion by including related terms and synonyms
- Three CVs: STW, subject terms, JEL codes
- No mapping yet, but goal is to normalize subject terms according to STW terms



Search in EconStor

telearbeit

Search

Extended Search Terms ▲

Telework Home-based work Virtual team Office technology

suggested terms:

- **Telework**, en
- **Telearbeit**, de
- telekommunikation / heimarbeit, de
- computerheimarbeit, de
- telework, en
- tele-working, de
- tele-heimarbeit, de
- distance working, en
- telecommuting, de
- telecommuting, en
- teleworking, en
- fernarbeit, de
- tele-arbeit, de
- teleheimarbeit, de
- telearbeit, de

results 1 to 10 of 10 total

Appears in

□ Tinbergen Institute Discussio Paper :

□ edition der Hans-Böckler-Stif :

□ IZA Discussion Papers :

□ Kiel Working Paper :

□ Ungeschützten Arbeitsverhältnisse: Frauen zwischen Risiko und neuer Lebensqualität :

scending

Teleworking and Congestion: A Dynamic Bottle
Gubins, Sergejs; Verhoef, Erik T.
2011
http://www.econstor.eu/bitstream/10419/87456
Gefunden in: Subject

Zusammen flexibel ist man weniger allein? Ein
Arbeitsform Coworking als Möglichkeit der Wis
Döring, Stefanie

t.borst@zbw.eu
http://zbw.eu/labs

Links

- https://wiki.surfnet.nl/display/standards/info-eu-repo
- http://www.daydream.co.uk/controlled_vocabulary.asp
- http://eprints.rclis.org/22447/
- https://www.liberquarterly.eu/articles/10.18352/lq.8035/
- https://www.jisc.ac.uk/guides/metadata/controlled-vocabulary
- http://connect.ala.org/node/209259
- http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/what.pdf
- https://wiki.duraspace.org/display/DSDOC5x/Authority+Control+of+Metadata+Values
- http://econstor.eu