

Automatic indexing of institutional repository content using SKOS

SWIB, 23.11.2020

Ricardo Eito-Brun
Universidad Carlos III
reito@uc3m.es

Index

- Lack of well-defined indexing practices is a common problem in most institutional repositories.
- Researchers typically assign keywords to submissions:
 - These terms are not extracted from a controlled vocabulary or thesaurus.
 - This leads to ambiguity and lack of specific indexing terms.
- Problem becomes clearer when aggregating content from multiple repositories.
- Approach: automatic assignment of descriptors taken from an existing thesaurus / KOS to already published contributions.
- The process is run with the help of a commercial tool, PoolParty.
- The experiment runs a process to automatically identify the “thesaurus concepts” that describe the content of the documents published in the institutional repository.

Previous work

- Previous work related to *collaborative, open innovation platforms*.
- Innovation is a knowledge-intensive process supported by “linkages”:
Services that connect the innovative organization with other entities in its context: universities, suppliers, clients, competitors, etc., and establish flows of knowledge and technology.
- OI platforms are web sites acting as “directories” of companies, people, etc., with the aim of support collaboration/innovation process keeping data about:
 - Researchers, lead users, university departments, research groups, small companies, etc.
 - Their work experience and technical achievements (patents, technical papers, product worksheets, etc.)
 - Innovation opportunities posted by different agents.

Previous work

- Purposes of OI platform:
 - Identify partners in a global context.
 - Get guidance to assess ideas sent in response to “innovation challenges”
 - Providing more information about solvers, to have a higher level of confidence on the proposed solutions.
- In that context, the need of better “matching capabilities” between innovation challenges and potential partners was identified.
- Potential use of specialized terminologies/vocabularies to describe entities, areas of expertise and achievements, challenges, etc.

Previous work

- Area: Biomedical engineering
 - Engineering discipline with several branches: bioinstrumentation, biomechanics, biomaterials, etc.
 - Main focus is on genetics, tissue engineering, medical software, simulators and imaging.
- The use of terminologies and controlled vocabularies help ensure the consistency of the descriptions and improve the capability to matching challenges with partners.
- MeSH was used to improve free text descriptions of agents, achievements and innovation opportunities using the *MeSH on demand* service.

Previous work

Indexing of “Entities”

- Companies, lead users, researchers, research groups, etc., who can contribute to the innovation process (as challengers or solvers).
- Attach data to entities: documents, patents, product descriptions, research projects.
- Data are indexed using MeSH on Demand.
- MeSH headings were added to free text descriptions.
- The “profile/record of the entity” is enhanced using the terms coming from their achievements (papers, CVs, etc.).

Previous work

Indexing of Innovation Opportunities

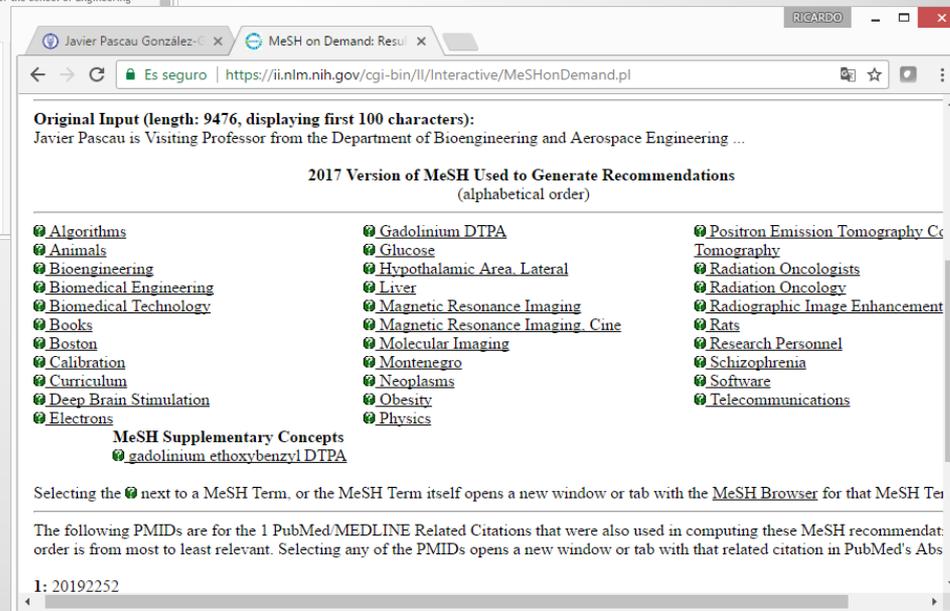
- **These are:** Challenges posted by any registered entity.
- **Functions:**
 - Posting opportunities.
 - Classify opportunities using MeSH.
 - Match the opportunity with the existing entities' profiles
 - Make a selective diffusion of information.

The Proposal Indexing of content with controlled terms

MeSH terms assigned to a researcher using his CV as an Entry (only titles of publications)



The screenshot shows the UC3M website profile for Javier Pascau González-Garzón. The header includes the UC3M logo and navigation tabs for ABOUT UC3M, STUDIES, ADMISSION, RESEARCH, and LIFE ON CAMPUS. The profile section includes a photo of the professor and a detailed biography: "Javier Pascau is Visiting Professor from the Department of Bioengineering and Aerospace Engineering at Universidad Carlos III de Madrid since 2011. He received his degree in Telecommunication Engineering from Universidad Politécnica de Madrid in 1999, a Master in Biomedical Technology and Instrumentation in 2005 and his PhD from Universidad Politécnica de Madrid in 2006. Before joining UC3M he was a research fellow at the Medical Imaging Lab in Hospital Gregorio Marañón in Madrid, where he also supervised the digital radiology migration project. His research lines include multimodal image quantification and registration both in..."



The screenshot shows the MeSH on Demand interface. The browser address bar indicates the URL: <https://i.nlm.nih.gov/cgi-bin/II/Interactive/MeSHonDemand.pl>. The page displays the following information:

- Original Input (length: 9476, displaying first 100 characters):**
Javier Pascau is Visiting Professor from the Department of Bioengineering and Aerospace Engineering ...
- 2017 Version of MeSH Used to Generate Recommendations**
(alphabetical order)

<input checked="" type="checkbox"/> Algorithms	<input checked="" type="checkbox"/> Gadolinium DTPA	<input checked="" type="checkbox"/> Positron Emission Tomography C
<input checked="" type="checkbox"/> Animals	<input checked="" type="checkbox"/> Glucose	<input checked="" type="checkbox"/> Tomography
<input checked="" type="checkbox"/> Bioengineering	<input checked="" type="checkbox"/> Hypothalamic Area, Lateral	<input checked="" type="checkbox"/> Radiation Oncologists
<input checked="" type="checkbox"/> Biomedical Engineering	<input checked="" type="checkbox"/> Liver	<input checked="" type="checkbox"/> Radiation Oncology
<input checked="" type="checkbox"/> Biomedical Technology	<input checked="" type="checkbox"/> Magnetic Resonance Imaging	<input checked="" type="checkbox"/> Radiographic Image Enhancement
<input checked="" type="checkbox"/> Books	<input checked="" type="checkbox"/> Magnetic Resonance Imaging, Cine	<input checked="" type="checkbox"/> Rats
<input checked="" type="checkbox"/> Boston	<input checked="" type="checkbox"/> Molecular Imaging	<input checked="" type="checkbox"/> Research Personnel
<input checked="" type="checkbox"/> Calibration	<input checked="" type="checkbox"/> Montenegro	<input checked="" type="checkbox"/> Schizophrenia
<input checked="" type="checkbox"/> Curriculum	<input checked="" type="checkbox"/> Neoplasms	<input checked="" type="checkbox"/> Software
<input checked="" type="checkbox"/> Deep Brain Stimulation	<input checked="" type="checkbox"/> Obesity	<input checked="" type="checkbox"/> Telecommunications
<input checked="" type="checkbox"/> Electrons	<input checked="" type="checkbox"/> Physics	

MeSH Supplementary Concepts
 gadolinium ethoxybenzyl DTPA

Selecting the next to a MeSH Term, or the MeSH Term itself opens a new window or tab with the [MeSH Browser](#) for that MeSH Term.

The following PMIDs are for the 1 PubMed/MEDLINE Related Citations that were also used in computing these MeSH recommendations. The order is from most to least relevant. Selecting any of the PMIDs opens a new window or tab with that related citation in PubMed's Abstracts.

1: 20192252

The Proposal Indexing of content with controlled terms

The image shows a screenshot of a web browser with two tabs. The first tab is titled 'GMV | Radiance | IORT' and displays a white paper titled 'RADIATION TREATMENT SIMULATION PLATFORM FOR IORT SUITABLE DEVICES'. The second tab is titled 'MeSH on Demand: Results' and shows a search interface for the 2017 version of MeSH. A callout box points to the search results, explaining the process of comparing MeSH terms with company classification in the BVJ database.

RADIATION TREATMENT SIMULATION PLATFORM FOR IORT SUITABLE DEVICES
TECHNOLOGICAL INNOVATION FOR IORT SUITABLE DEVICES TREATMENT SIMULATION

Product white paper: Compare terms from MeSH with Company classification in BVJ database: "Computer programming services", "Engineering services" (SIC, CAE, NAICS, CNAE...)

2017 Version of MeSH Used to Generate Recommendations
(alphabetical order)

<input type="checkbox"/> Achievement	<input type="checkbox"/> Hospitals, University	<input type="checkbox"/> Patient Care Team	<input type="checkbox"/> Modulated
<input type="checkbox"/> Algorithms	<input type="checkbox"/> Humans	<input type="checkbox"/> Photons	<input type="checkbox"/> Radius
<input type="checkbox"/> Animals	<input type="checkbox"/> Incidence	<input type="checkbox"/> Physics	<input type="checkbox"/> Rectal Neoplasms
<input type="checkbox"/> Brachytherapy	<input type="checkbox"/> Inventions	<input type="checkbox"/> Positron Emission Tomography	<input type="checkbox"/> Recurrence
<input type="checkbox"/> Breast Neoplasms	<input type="checkbox"/> Investments	<input type="checkbox"/> Computed Tomography	<input type="checkbox"/> Sarcoma, Ewing
<input type="checkbox"/> Calibration	<input type="checkbox"/> Magnetic Resonance Imaging	<input type="checkbox"/> Probability	<input type="checkbox"/> Software
<input type="checkbox"/> Cell Survival	<input type="checkbox"/> Monte Carlo Method	<input type="checkbox"/> Prone Position	<input type="checkbox"/> Specialization
<input type="checkbox"/> Certification	<input type="checkbox"/> Needles	<input type="checkbox"/> Quality Control	<input type="checkbox"/> Surgeons
<input type="checkbox"/> Computer Systems	<input type="checkbox"/> Ointments	<input type="checkbox"/> Radiation Oncologists	<input type="checkbox"/> Tail
<input type="checkbox"/> Documentation	<input type="checkbox"/> Organs at Risk	<input type="checkbox"/> Radiation Oncology	<input type="checkbox"/> Tumor Burden
<input type="checkbox"/> Electronics	<input type="checkbox"/> Ovary	<input type="checkbox"/> Radiobiology	<input type="checkbox"/> Water
<input type="checkbox"/> Female	<input type="checkbox"/> Paintings	<input type="checkbox"/> Radiosurgery	<input type="checkbox"/> Workflow
<input type="checkbox"/> Goals	<input type="checkbox"/> Particle Accelerators	<input type="checkbox"/> Radiotherapy, Intensity-	<input type="checkbox"/> X-Rays

Selecting the next to a MeSH Term, or the MeSH Term itself opens a new window or tab with the [MeSH Browser](#) for that MeSH T

The following PMIDs are for the 10 PubMed/MEDLINE Related Citations that were also used in computing these MeSH recommen

The Proposal Indexing of content with controlled terms

Cleveland Clinic: Novel Implant for the Treatment of Pelvic Organ Prolapse

TAGS: Cleveland Clinic Life Sciences Ideation

STATUS: Awarded
ACTIVE SOLVERS: 91
POSTED: 18 nov 2014
SOURCE: InnoCentive

Pelvic organ prolapse (POP) is the descent of pelvic organs including the bladder, uterus surrounding structural supports. Recent studies reveal that women have a 12.6% risk of u over their lifetime. Operations to treat POP include native tissue procedures which use or are augmented with graft. The goal of POP surgical repair is to provide long-lasting impro function while minimizing morbidity and recurrence. However, up to 6% of women who he repair will require reoperation at 5 years for mesh complications and approximately 10% prolapse. The development of a better implant could accomplish this goal.

The Seeker is interested in a graft material the use of which will improve outcomes of PO biomechanically complementary material ideal for retaining or returning pelvic organ and bearing state. This material should also be easily implanted surgically and ideally would of tissue ingrowth had occurred.

This is an Ideation Challenge with a guaranteed award for at least one submitted solution

PREMIUM CHALLENGE

Challenge: Compare tags give in the Innocentive site with terms from MeSH

Original Input (length: 4824, displaying first 100 characters):
Cleveland Clinic: Novel Implant for the Treatment of Pelvic Organ Prolapse ;
TAGS: Cleveland Clinic ...

2017 Version of MeSH Used to Generate Recommendations
(alphabetical order)

<input type="checkbox"/> Absorbable Implants	<input type="checkbox"/> Goals	<input type="checkbox"/> Pelvic Organ Prolapse	<input type="checkbox"/> Research Personnel
<input type="checkbox"/> Atrophy	<input type="checkbox"/> Humans	<input type="checkbox"/> Reconstructive Surgical Procedures	<input type="checkbox"/> Urinary Bladder
<input type="checkbox"/> Awards and Prizes	<input type="checkbox"/> Intellectual Property	<input type="checkbox"/> Rectum	<input type="checkbox"/> Urinary Incontinence
<input type="checkbox"/> Biological Science Disciplines	<input type="checkbox"/> Medicine	<input type="checkbox"/> Reoperation	<input type="checkbox"/> Uterus
<input type="checkbox"/> Fellowships and Scholarships	<input type="checkbox"/> Patient Care	<input type="checkbox"/> Research	
<input type="checkbox"/> Female	<input type="checkbox"/> Pelvic Floor Disorders		

Selecting the next to a MeSH Term, or the MeSH Term itself opens a new window or tab with the [MeSH Browser](#) for that MeSH T

The following PMIDs are for the 10 PubMed/MEDLINE Related Citations that were also used in computing these MeSH recommend order is from most to least relevant. Selecting any of the PMIDs opens a new window or tab with that related citation in PubMed's AT

1: 19094645	3: 21126129	5: 24118430	7: 12619181	9: 25656453
2: 21592280	4: 23939382	6: 22326614	8: 24866279	10: 22857745

Disclaimer: These MeSH Terms are machine generated by MTI and **DO NOT reflect any human review**. MTI may recommend MeS not explicitly found in the text and may not recommend MeSH Terms that are in the text. This is a result of machine logic that attempt emulate human indexer behavior in characterizing biomedically relevant parts of the text. These results will undoubtedly differ from

What happens with Open, Institutional Repositories?

- Contributions/Items typically include keywords assigned by the authors / self-archiving policies:
 - No terminology / vocabulary control at all.
 - Sometimes, librarians assign general, high-level categories to items.
- But, the benefits of controlled vocabularies are missing:
 - Users cannot be led to relevant content (more specific, generic or related).
 - Users cannot explore/browse the repository from an conceptual perspective.

What happens with Open, Institutional Repositories?

- Huge problem when integrating items from different repositories
 - Syntactic interoperability has been achieved with OAI-PMH and similar protocols.
 - We are far-away of semantic interoperability.
 - Sites like Recolecta that collect items from different repositories.
 - User provided keywords are not enough to ensure accurate indexing/retrieval of content.
 - At “aggregator sites”, typical distinction between communities and collections makes no meaning
 - We can rely uniquely in the free-text abstracts and keywords provided by authors, plenty of issues.

Experiment

- Experiment was conducted with metadata / items published in the institutional repository of Universidad Carlos III de Madrid: e-Archivo.
- With the aim to assess the potential use of *automatic indexing tools* to assign descriptors coming from a recognised, widely used thesaurus.
- Tool selected was PoolParty, which provides the capability of matching textual descriptions with terms defined in SKOS thesauri.
- Experiment conducted with a subset of 6000 records.

Experiment

- Universidad Carlos III joined Open Access declaration on 12.06.2006 (see Max Planck) being the 172 university out of 648 supporting that initiative:
<https://openaccess.mpg.de/319790/Signatories>
- E-Archivo was created in November 2006 as part of a larger initiative with the purpose of setting up:
 - An open archive network public universities in Madrid.
 - “e-Science” web portal giving centralized, open access to content published by members of the consortia.

Experiment

- E-Archive main objectives include:
 - Integrate and preserve the intellectual production of the University.
 - Increment the visibility of the works, authors, and the University itself.
 - Increment the impact of the scientific output.
 - Give free Access to that content.
- Repository is based on DSpace.
- Voluntary, self-archiving submission of documents by authors, with the exception of:
 - Research funded with public budget (Ley 14/2011, de la Ciencia, la Tecnología y la Innovación, art. 37).
 - Doctoral dissertations (Real Decreto 99/2011, art. 14.5).

Why UNESCO Thesaurus?

- Survey conducted among Spanish University Libraries.
- To get knowledge on the use of controlled vocabularies on their institutional repositories.
- Contact details identified through CRUE (47 universities).
- Identification of metadata used to encode “keywords” and “descriptors”.
- Thesauri/vocabularies in place.
- 50% using CV chose UNESCO Thesaurus.



Why UNESCO Thesarus?

ESTUDIO VOCABULARIOS CONTROLADOS EN REPOSITORIOS INSTITUCIONALES DE UNIVERSIDADES PÚBLICAS ESPAÑOLAS				
Universidad	Nombre repositorio	Software	Elemento Dublin Core	Vocabulario controlado
1. Univ. de Alcalá de H.	e-Buah	DSpace	dc.subject.other y dc.subject.eciencia (materia)	NO
2. Univ. de Alicante	RUA	DSpace	dc.subject y dc.subject.other	NO
3. Univ. de Almería	riUAL	DSpace	dc.subject	NO
4. Univ. de Barcelona	Dipòsit Digital	DSpace	dc.subject.classification y dc.subject.other	SÍ
5. Univ. de Burgos	RIUBU	DSpace	dc.subject / dc.subject.other y dc.subject.unesco	SÍ
6. Univ. de Cádiz	RODIN	DSpace	dc.subject	NO
7. Univ. de Cantabria	UCrea	DSpace	dc.subject.other	NO
8. Univ. de Castilla La Mancha	RUIdeRA	DSpace	dc.subject	NO
9. Univ. de Córdoba	HELVIA	DSpace	dc.subject	NO
10. Univ. de Extremadura	DEHESA	DSpace	dc.subject y dc.subject.unesco	SÍ (materias UNESCO)
11. Univ. de Girona	DUGiDocs	DSpace	dc.subject	NO
12. Univ. de Granada	DIGIBUG	DSpace	dc.subject	NO
13. Univ. de Huelva	Arias Montano	DSpace	dc.subject.other y dc.subject.unesco	SÍ (materias UNESCO)
14. Univ. de las Islas Baleares	UIBrepositori	DSpace	dc.subject.other / dc.subject.classification y dc.subject.keywords	SÍ
15. Univ. de Jaén	RUJA	DSpace	dc.subject.other	NO
16. Univ. de La Coruña	RUC	DSpace	dc.subject	NO
17. Univ. de La Laguna	RIULL	DSpace	dc.subject.keyword	NO
18. Univ. de Las Palmas de Gran Canaria	ACCEDA	DSpace	dc.subject y dc.subject.other	SÍ
19. Univ. de León	BULERIA	DSpace	dc.subject.other y dc.subject.unesco	SÍ
20. Univ. de Lérida	Repositori Obert UdL	DSpace	dc.subject	NO
21. Univ. de Málaga	RIUMA	DSpace	dc.subject y dc.subject.other	NO
22. Univ. de Murcia	DIGITUM	DSpace	dc.subject y dc.subject.other	SÍ (CDU)
23. Univ. de Oviedo	RUO	DSpace	dc.subject	NO
24. Univ. del País Vasco	ADDI	DSpace	dc.subject y dc.subject.categoria	NO* (listado de materias)
25. Univ. de Salamanca	GREDOS	DSpace	dc.subject y dc.subject.unesco	SÍ (materias UNESCO)
26. Univ. de S. de Compostela	MINERVA	DSpace	dc.subject	NO

Why UNESCO Thesarus?

27. Univ. de Sevilla	idUS	DSpace	dc.subject	NO
28. Univ. de Valencia	RODERIC	DSpace	dc.subject y dc.subject.unesco	Sí (materias UNESCO)
29. Univ. de Valladolid	UVaDOC	DSpace	dc.subject.classification	NO
30. Univ. de Vigo	Investigo	DSpace	dc.subject.unesco	Sí (materias UNESCO)
31. Univ. de Zaragoza	ZAGUAN	CDS Invenio	-	NO
32. Univ. Autónoma de Barcelona	DDD	CDS Invenio	dc.subject	NO
33. Univ. Autónoma de Madrid	Biblos e-archivo	DSpace	dc.subject.other y dc.subject.eciencia (materia)	NO* (listado de materias)
34. Univ. Carlos III de Madrid	e-archivo	DSpace	dc.subject.other y dc.subject.eciencia (materia)	NO* (listado de materias)
35. Univ. Complutense de Madrid	e-prints Complutense	E-Prints	dc.subject	NO* (listado de materias)
36. Univ. Internacional Andalucía	UNIA	DSpace	dc.subject	NO
37. Univ. Jaume I	Repositori UJI	DSpace	dc.subject	NO
38. Univ. Miguel Hernández	RediUMH	DSpace	dc.subject y dc.subject.other (materia)	Sí (CDU)
39. Univ. Pablo de Olavide	RIO	DSpace	dc.subject	NO
40. Univ. Politécnica de Cataluña	UPCommons	DSpace	dc.subject / dc.subject.other / dc.subject.lcsh y dc.subject.lemac	Sí
41. U. Politécnica de Cartagena	UPCT	DSpace	dc.subject / dc.subject.other y dc.subject.unesco	Sí (materias UNESCO)
42. Univ. Politécnica de Madrid	Archivo Digital UPM	E-Prints	-	NO* (listado de materias)
43. Univ. Politécnica de Valencia	RiuNet	DSpace	dc.subject y dc.subject.classification	NO
44. Univ. Pompeu Fabra	e-repositori	DSpace	dc.subject.keyword	NO
45. Univ. Pública de Navarra	Academica-e	DSpace	dc.subject	NO
46. Univ. Rey Juan Carlos I	BURJC Digital	DSpace	dc.subject y dc.subject.unesco	Sí (materias UNESCO)
47. Univ. Rovira i Virgili	R. institucional	Fedora	-	NO

Why UNESCO Thesaurus?

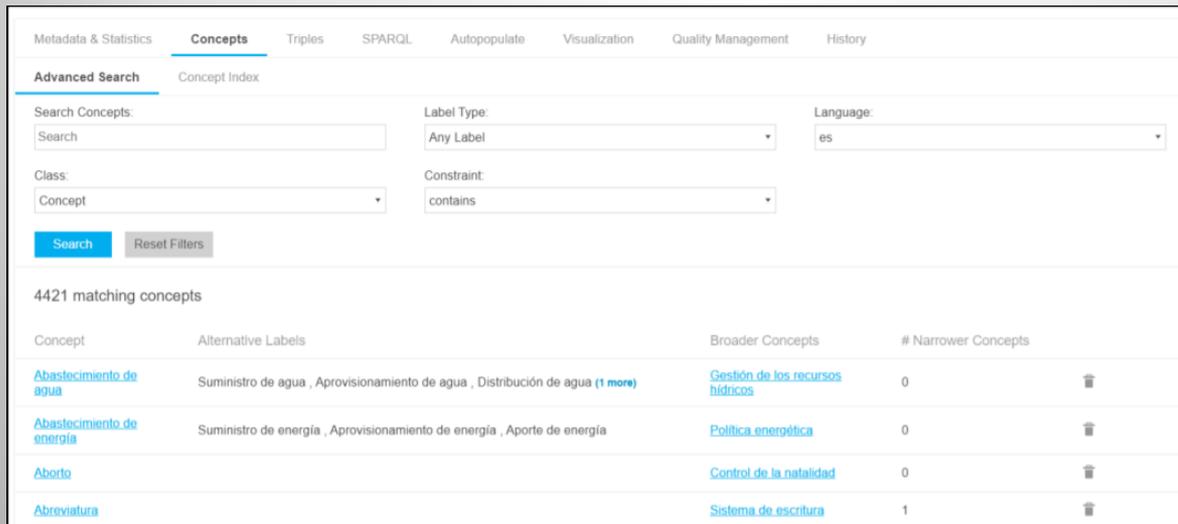
- Question: why did you select UNESCO Thesauri?
 - Relevance of dissertations
 - Already used by other libraries / repositories.
 - Simple, easy to use schema.
 - Vocabulary provides specific terms to Support “detailed indexing”.
 - Support to interoperability.
- Universities using controlled vocabularies stated that:
 - People in charge of indexing items are Library staff (46%) or staff dedicated to the repository (27%).
 - In other cases, it was made by reviewers in charge of approval.

Why UNESCO Thesaurus?

- Question: % of Thesaurus term being used?
 - Less than 40%
 - Significant differences between universities (área to research)
 - Most of the repositories (82%) calculates the documents per term/category.
 - Search/Query logs are not used in most of the cases.

Experimentation

- Downloaded records were processed with PoolParty tool.
- Before processing the OAI-PMH downloaded records, the UNESCO Thesaurus was incorporated to PoolParty.
- It includes 4421 “concepts” with different “linguistic representations”.



The screenshot displays the PoolParty web interface. At the top, there are navigation tabs: Metadata & Statistics, Concepts (selected), Triples, SPARQL, Autopopulate, Visualization, Quality Management, and History. Below the tabs, there are two sub-tabs: Advanced Search (selected) and Concept Index. The search form includes fields for Search Concepts (with a search input), Label Type (set to Any Label), Language (set to es), Class (set to Concept), and Constraint (set to contains). There are Search and Reset Filters buttons. Below the search form, it indicates 4421 matching concepts. A table lists the results with columns for Concept, Alternative Labels, Broader Concepts, and # Narrower Concepts.

Concept	Alternative Labels	Broader Concepts	# Narrower Concepts
Abastecimiento de agua	Suministro de agua , Aprovechamiento de agua , Distribución de agua (1 more)	Gestión de los recursos hídricos	0
Abastecimiento de energía	Suministro de energía , Aprovechamiento de energía , Aporte de energía	Política energética	0
Aborto		Control de la natalidad	0
Abreviatura		Sistema de escritura	1

Experimentation

- The tool provides a “corpus management” feature to analyse / match textual descriptions with CV terms.
- For the initial set of documents (400), 224 terms from the Thesaurus were “matched”, 643 for the whole set.
- Average of 4 terms per document.

The screenshot displays the docsOAI Corpus Search interface. At the top, there is a navigation bar with tabs for 'Metadata & Statistics', 'Extracted Concepts', 'Extracted Terms', and 'Corpus Documents'. The 'Metadata & Statistics' tab is active. Below the navigation bar, there are two main sections: 'Corpus Analysis Summary' and 'Corpus Summary'.

Corpus Analysis Summary

Complete	
Last Calculation	15.06.2020 - 17.04
Extracted Concepts	224
Extracted Terms	83,738
Concept Occurrences	1,583
Term Occurrences	162,124
The quality of the suggested terms may be low. Add more documents to the corpus to improve quality of terms.	
Poor	

Corpus Summary

Stored Documents	308
Overall Filesize	0.84 MB
Language	es
Created by	ricardo_eitobrun
Created	15.06.2020 - 16.59
Last Modified	15.06.2020 - 17.02
Server	Local

Experimentation

- The tool provides different views on the most-used concepts sorted by frequency and identify the items they are assigned to.
- Export capabilities gives the choice of reusing the assignment of terms to “re-index” ítems (DSpace massive updates capabilities, now in the process).

Extracted Concepts					
Preferred Label	Frequency	Relevance	Most Frequent Label	Broader Concepts	Concept Scheme
España	130	7.67	españa	País mediterráneo , País de la OCDE , País de la CEE , Europa Occidental , Unión Europea	Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Gobierno	124	7.73	gobierno		Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Ensayo	74	3.08	control	Método científico	Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Sistema económico	68	3.19	economía		Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Universidad	58	0.38	universidad	Instituto de enseñanza superior	Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Internet	58	2.89	internet	Red informática	Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Comunidad	52	3.85	comunidad	Asentamiento humano	Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Casa	38	2.78	departamento	Vivienda	Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Robot industrial	34	4	robot	Máquina	Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>
Informática	33	0.42	informática		Tesouro de la UNESCO <input checked="" type="checkbox"/> <input checked="" type="checkbox"/>

1 2 3 4 5 6 7 8 9 10 Next Last

Conclusions

- Constraints:
 - Experiment was completed on a limited set of records
 - Experiment conducted on metadata (full-text processing was discarded due to technical constraints, although it would be possible)
 - Possibility of checking additional tools.
- In any case, the combined use of SKOS and automatic term extraction result in positive outcomes:
 - Identification of terms defined in KOS that are used in the documents.
 - Automatically assignment of terms from KOS, with no effort from staff.
 - Using KOS opens the possibility of implementing better “browsing capabilities” on repository contents, even if this content has not been indexed before using that vocabulary.

Conclusions

- But most of the benefits can be obtained when thinking on integration / aggregation of metadata.
- Once thesaurus descriptors are assigned to the existing documents:
 - Establish cross-searching methods to enhance search through repositories (index each site separately).
 - Enrich aggregated items' descriptions with descriptors coming from KOS, improving content classification at aggregators' sites (index aggregated metadata).
- In any case, the semantic integration of content and metadata from different sources can be improved by the use of common indexing languages.

Questions.

- Thanks!