

A EU Workflow for Seamless Maintenance and Publication of Data, Metadata and Legal Acts

Armando Stellato^{1,2}, Manuel Fiorelli^{1,2} Andrea Turbati² and Tiziano Lorenzetti²

¹ University of Rome Tor Vergata, Dept of Enterprise Engineering, Via del Politecnico 1, Rome, Italy

² Lore Star srl, via Leonida Rech 77, Rome, Italy

Abstract

In 2016, we began collaborating with the Publications Office of the EU on the development of a platform for collaborative management of Linked Open Datasets. The platform, VocBench 3, assumed soon a key role in the development and maintenance of all metadata resources adopted by the office, while becoming one of the most popular platforms for management of semantic resources worldwide. After VocBench was already in production within the office, a second platform, ShowVoc, addressed the needs of publishing the data in most convenient ways to the public. More recently, we have developed LegalHTML, a language for the representation of LegalActs, binding representation needs, semantics and structure in a single format. Nowadays, with the Publications Office going to adopt LegalHTML, the circle is closed: LegalHTML is general and flexible enough to support diverse legal traditions yet can be purposed to specific ones by extending its semantic model with specific ontologies and reference datasets. Adoption of the EU ELI model and exploitation of the several authority resources from the Publications Office constitute a strong semantic backbone which in the future will enhance data and document retrieval and sharing.

Keywords

Publication Workflow, Linked Open Data, Legal Documents, Semantic Standards

1. Introduction

In 2016, the ISA2 programme² of the EU for “Interoperability solutions for public administrations, businesses and citizens” funded the realization of a platform for the collaborative development and maintenance of semantic resources. Born on the legacy of two previous experiences [1] matured within European Research Frameworks and, later, in a collaboration between the Food and Agriculture Organization (FAO) of the United Nations and the University of Rome Tor Vergata, VocBench 3 [2] – such is the name of the realized platform – immediately gathered consensus around its unique combination of editing features, support

for collaborative development and wide support of Semantic Web standards, all enclosed in an open source, free-to-use, platform.

VocBench 3 (or, shortly, VB) is, today a fully-fledged collaborative environment with dedicated support for developing a wide variety of semantic resources (e.g. ontologies, knowledge organization systems, OntoLex lexicons, EDOAL alignments and RDF datasets in general). Besides its openness and availability to the masses, VocBench has a central role in the European Commission as an instance of the platform, hosted at the Publications Office of the EU (in short, OP), is currently managing hundreds of datasets developed by the OP itself or by many other directorate generals of the EU commission.

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29-31, 2023, Pisa, Italy

EMAIL: stellato@uniroma2.it (A. 1); manuel.fiorelli@uniroma2.it (A. 2); andrea.turbati@lorestar.it (A. 3) tiziano.lorenzetti@lorestar.it (A. 4).

ORCID: 0000-0001-5374-2807 (A. 1); 0000-0001-7079-8941 (A. 2); 0000-0002-6214-4099 (A. 3); 0000-0001-5676-8877 (A. 4)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

² <https://ec.europa.eu/isa2/>

Later on, in the context of another action of the ISA2 programme, PMKI (Public Multilingual Knowledge Infrastructure) and, in a sense, from a rib of VocBench, we realized a companion system to it focused on data publication and consumption. ShowVoc, this is the name of the system, complements VB's editing capabilities by providing a platform supporting Linked Open Data portals, showing a streamlined interface for data browsing targeted at end users, advanced search capabilities across multiple datasets, integration inside existing portals and support for http resolution and content negotiation.

Finally, in early 2022, we had the occasion to develop, in two projects directly funded by the OP, a language for the representation of legal acts, LegalHTML, with the mission to overcome the limitations of current approaches (mostly based on XML and, in the best case, indirect use of ontologies) and unifying, in a single language: representation, structure and semantics. This experience gave us the occasion to “close the circle” on showing the usefulness of metadata, as LegalHTML is fully exploiting the range of ontologies, authority tables and other semantic resources in force at the Publications Office.

In this paper, we give a short overview of these solutions in the context of their applications at the Publications Office of the EU and, concerning the first two, for the wide adoption they have met both within EU soil and in extra-EU countries.

2. The VocBench Project

VocBench (a screenshot available in Figure 1), yet from its conception, embodying the principles of openness (in terms of availability of both applications and their source code) and participation of the ISA² programme, was born on very clear requirements provided by a community of users already acquainted with Semantic Web standards, Linked Open Data and knowledge management. A first meeting was held in early 2016 at the Publications Office, gathering stakeholders (mostly from the public sector) and consultants to line up the set of requirements for the system. This is the list that emerged:

- R1. Multilingualism
- R2. Controlled Collaboration
- R3. Data Interoperability and Integrity
- R4. Software Interoperability and Extensibility
- R5. Data Scalability

- R6. Under-the-hood data access and modification
- R7. Adaptive Context and Ease-of-use
- R8. RDF Languages Support
- R9. Maintainability (Architecture and Code Scalability)
- R10. Full Editing Capability (RDF Observability and Reachability)
- R11. Provenance
- R12. Versioning Support
- R13. Dataset-level Metadata Descriptions
- R14. Customizable User Interface
- R15. Everything's RDF

Indeed, the first 7 requirements were inherited from the first two incarnations of the system, developed within the aforementioned collaboration between FAO and the University of Rome Tor Vergata, and reflect the needs of a scenario with world-wide contributions to a large shared resource such as the Agrovoc thesaurus [2,3] of FAO. No wonder, multilingualism was a must and requirement n° 1 as much as it is in the multicultural and multilingual European scenario, R2 reflects the exigencies of a large organization with well-defined policies for maintaining open but yet authoritative resources, R3,4 embody the modern approach of an open system for realizing open shareable data while R5 concerns with the non-trivial size of large general-purpose vocabularies, terminologies and thesauri.

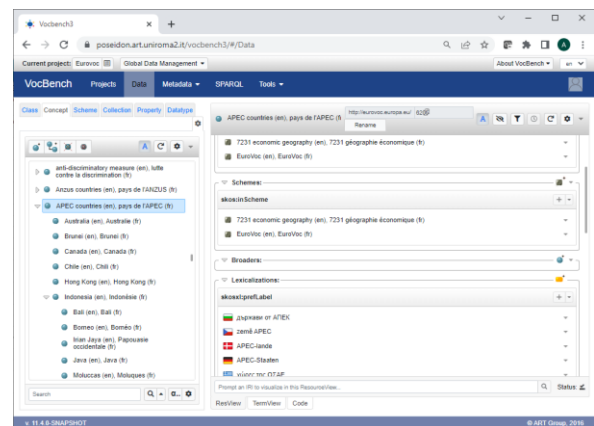


Figure 1: VocBench, editing the EuroVoc thesaurus of the EU

Requirements 6,7 and the following ones added during the stakeholders' meeting show more elaborate, technical, desiderata for a system that had to be natively adherent to semantic standards and ready to cover all aspects of the production of a resource from its initial conceptualization to its publication on the Web,

while matching the diverse policies of organizations, industries and P.A.s for the maintenance of semantic resources.

Nowadays, VocBench sets an unprecedented milestone in the management of semantic resources being the only free & open-source (and, to our knowledge, possibly the same even if not restricting to this category) solution offering collaborative management services with extensive, dedicated, support for a broad range of semantic models and, thus, type of resources.

3. ShowVoc

ShowVoc (here below in Figure 2, SV. shortly) aims at the development of open data portals focusing on terminological and linguistic content.

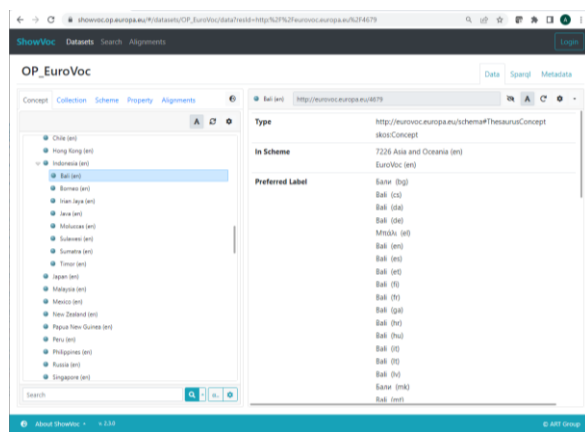


Figure 2: ShowVoc, browsing EuroVoc

The portal is a sort of read-only VB, including the various tree panels, the graph views, the resource view (i.e. the panel where the details of each selected resource are shown) with a focus on efficiency and streamlined fruition of content.

ShowVoc inherits most of the requirements of VB besides those related to editing, while introducing new ones specific to its mission:

- SV-R1. Represent datasets-as-a-whole as a further first-class-citizen resource
- SV-R2. Offer cross-dataset services

These services and views have been translated into dedicated capabilities such as the possibility to browse linksets between datasets (see Figure 3), to list the triples associated with these linksets, to perform efficient cross-project search, to exploit multilingual resources (or, in the near future, the

links among these resources) to provide authoritative translations and a metadata page with both qualitative and quantitative information on the hosted datasets adopting the most common vocabularies for metadata representation (i.e. VoID [4], LIME [5], DCAT [6] – its third version, still under development – and a dedicated extension wrapping these vocabularies and binding them under a common umbrella, developed specifically for ShowVoc).

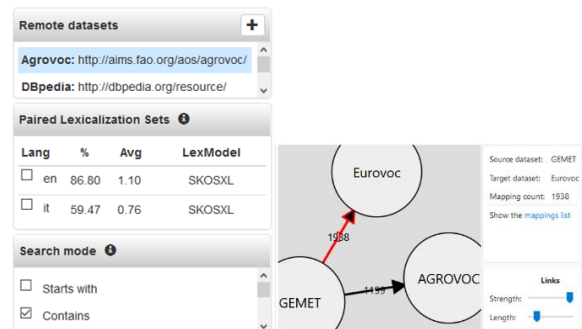


Figure 3: Assisted search (on the left) and browsing linksets in ShowVoc (on the right)

ShowVoc is thus analogous to data repository software, such as CKAN³ (used in turn by many public data portals, such as US Gov⁴, or EU Data Portal⁵), Invenio⁶ (used by Zenodo⁷ and CERN Open Data⁸) and OntoPortal⁹ (based on BioPortal [7]), yet it offers a unique combination of content fruition services, metadata storage and derived application services with respect to its counterparts, making it a convenient choice if the focus is specifically on RDF resources.

Its specialization on RDF resources drove a further important feature aimed at simplifying the work of users needing to publish them: support for http resolution and content negotiation. Setting up all the required machinery for implementing Linked Open Data publication policies and best practices is a non-trivial and time consuming task. ShowVoc simplifies this task by acting as a sort of orchestrator for the whole process and by providing ready-to-use implementations for most of the task configuration. Aspects such as content negotiation based on the requested formats and provision of both RDF content (generated by the same system) and HTML one (which can be the UI of ShowVoc itself or an external page) are all managed through a centralized dashboard allowing for the configuration of each dataset.

³ <https://ckan.org/>

⁴ <https://data.gov>

⁵ <https://data.europa.eu/>

⁶ <https://inveniosoftware.org/>

⁷ <https://zenodo.org/>

⁸ <https://opendata.cern.ch/>

⁹ <https://ontoportal.org/>

Furthermore, the ability to embed its user interface within other pages makes it a convenient solution for other portals that can thus focus more on the overall offer and spare the non-trivial implementation of browsing semantic resources.

4. LegalHTML

In early 2022, we conducted a study, funded by the Publications Office of the European Union, exploring the feasibility of a solution for the publication of legal acts that would streamline the complex publication workflow of legal acts implemented at the OP. The workflow currently includes several steps: drafting, proof-reading, finalization and production of several versions of the same acts purposed for different loci, such as official journal publication, semantic indexing, dissemination, etc.. Even the first steps, while being linear, imply the realization of progressive iterations of an act through different formats, as these are more or less convenient depending on the task to be performed, on the (usual, expected) background of the staff working on it and on other boundary conditions. For instance, the initial drafting and proofreading are performed on LegisWrite, a format – developed within the OP itself – consisting in a series of styles for Microsoft Word, adopted in a rigorous manner to annotate the different parts of an act. The drafting is supported by a plugin for Word, which proposes styles through dedicated menus and buttons. This method is highly prone to human error as there is no enforcing of the proper styles nor a guided assistance for the user but keeps being the preferred method due to the WYSIWYG nature and the popularity of the editor. Subsequent steps include transformations and editing of the input into progressively more elaborated models, such as the Common Vocabulary of the Interinstitutional Metadata and Formats Committee¹⁰ (COV, in short), or the Formalized Exchange of Electronic Publications¹¹ (Formex), lightweight (relatively speaking, considering that they are about legal acts) XML formats mainly focused on the structure of legal acts and, to a limited extent, to semantics. Recently, the OP decided to embrace the popular OASIS standard Akoma Ntoso [8], even though in a mutated version, tailored for the legal tradition of the EU, called AKN4EU. This format, building up on the already established standard, is being adopted for

final representation of the semantic content of legal acts. Finally, content presentation is undoubtedly a necessity for human consumption and, alongside the PDF versions of the original acts from the Official Journal, the OP publishes on its EUR-Lex¹² site, in HTML format, convenient multilingual versions of acts, that are also “consolidated” according to the evolutions that each single regulation, directive, etc.. undergoes following amendments and other modifications in time. EUR-Lex completes the publication of the documents with rich metadata, in RDF, stored as RDFa. The metadata is not stored within the single HTML documents representing different consolidations and language translations of acts, rather in an entry page that binds them all. RDFa is in fact not used as in its typical purpose of an “annotation” language, as no text content is being annotated, but as a convenient, standard, vector for storing these metadata triples. Indeed, the RDFa is stored in repeated <meta> elements (each tag contains a single triple) of the said entry page. For the representation of metadata, a common metadata model is being adopted, represented by the ELI (European Legislation Identifier) ontology¹³.

With the intent of covering the aforementioned steps of the workflow through a single, coordinated, solution, we have realized an HTML-based language (i.e. exploiting recent HTML support for extensions, thus realizing a language of its own, targeted at a specific domain) for representing legal content, dubbed LegalHTML, featuring all structural aspects of an act, such as articles, paragraphs, items, references, etc.. supporting a semantically-rich representation of all such elements and references to entities of the legal domain. Furthermore, LegalHTML addresses consolidation of an act and its subsequent modifications into a single document using an efficient tree-based model. Finally, we imbued the document model with API supporting rendering and point-in-time visualization of legal acts (see Figure 4). Metadata, consolidation information and other relevant information are represented by a dedicated ontology. In fact, other ontologies and controlled vocabularies can be combined to ground LegalHTML in different legal traditions (i.e., different ways to represent laws in different countries), without violating the integrity and generality of the language.

¹⁰ <https://op.europa.eu/it/web/eu-vocabularies/cov>

¹¹ <https://op.europa.eu/it/web/eu-vocabularies/formex>

¹² <https://eur-lex.europa.eu/>

¹³ <https://op.europa.eu/web/eu-vocabularies/eli>

Table of contents	COMMISSION DECISION
Top	of 12 June 2008
Article 1 - Subject-matter	establishing a specific control and inspection programme related to the salmon and cod stocks in the Baltic Sea
Article 2 - Scope	
Article 3 - Definitions	(notified under document number C(2008) 2558)
Article 4 - Commission inspections	(2008/589/EC)
Article 5 - Member State inspections	THE COMMISSION OF THE EUROPEAN COMMUNITIES,
Article 6 - Joint inspection and surveillance activities	Having regard to the Treaty establishing the European Community ,
Article 7 - Information	Having regard to Council Regulation (EEC) No 2847/93 of 12 October 1993 establishing a control system applicable to the common fisheries policy ⁽¹⁾ , in particular Article 34c(1) thereof,
Article 8 - Evaluation	Whereas:
Article 9 - Addresses	(1) Council Regulation (EC) No 1098/2007 establishing a multi-annual plan for the cod stocks in the Baltic Sea and the fisheries exploiting those stocks, lays down the conditions for the sustainable exploitation of cod in the Baltic Sea and the rules on monitoring, control and surveillance of such activities.
ANNEX I	(2) Council Regulation (EC) No 2371/2002 of 20 December 2002 on the conservation and sustainable exploitation of fisheries resources under the Common Fisheries Policy ⁽²⁾ provides for control activities by the Commission and cooperation between Member States to ensure compliance with the rules of the Common Fisheries Policy.
ANNEX II	(3) To ensure the success of the multi-annual plan for the cod stocks in the Baltic Sea and the fisheries exploiting those stocks, it is necessary to establish a specific control and inspection programme.
Versions	(4) The specific control and inspection programme should be established for a period of three years. The results obtained by the application of the specific control and inspection programme should be periodically evaluated by the Member States concerned in cooperation with the Community Fisheries Control Agency (CFCA) set up by Council Regulation (EC) No 768/2005 ⁽³⁾
10/8/2012	(5) Cooperation between Member States concerned should be encouraged so as to enhance uniformity of inspection and surveillance practices and help develop the coordination of control activities between the competent authorities of those Member States.
18/5/2012	(6) Joint inspection and surveillance activities should be carried out in accordance with joint deployment plans established by the CFCA.
19/2/2011	(7) The measures provided for in this Decision have been established in concert with the Member States concerned.
Legal act	(8) The measures provided for in this Decision are in accordance with the opinion of the Management Committee for Fisheries and Aquaculture.
	HAS ADOPTED THIS DECISION:
	<i>Article 1</i>
	Subject-matter
	This Decision establishes a specific control and inspection programme to ensure:
	<small>link to the harmonised implementation of the multilingual data set on the Regulation (EC) No 1098/2007 for cod</small>

Figure 4: a decision of the EC in LegalHTML. The “Versions” menu consolidates the document

Finally, another relevant aspect is that all the metadata information is conveniently stored within the document. While not being prescriptive (indeed any standard method can be adopted), the suggested solution is to store metadata that is not directly annotating textual content in a `<script>` element, using any RDF serialization (preferably Turtle) and to use RDFa to semantically annotate the textual content of the act. LegalHTML has a core ontology for providing core semantics and references, while offers explicit extension points for operating within specific legal traditions. In the case of the EU, the already mentioned ELI ontology has been adopted while the whole range of semantic authority resources published by the Publications Office, maintained through VB and published (also) through SV has been adopted.

The outcome of the study has been accepted and a second project funded the finalization of the specifications and the production of a sample document set that could help contractors in realizing LegalHTML versions of the documents. The project has recently concluded and LegalHTML is now scheduled for production.

5. Adoption

While LegalHTML is a too recent achievement to see any concrete diffusion beyond

its own crib, VB and SV have something to say. A heterogeneous user community has grown in these years around VB, including large organizations and the public sector in general, companies needing VB as users and companies featuring it as their platform of choice in their range of offered RDF services, consultants needing a modeling environment, etc.

The Publications Office of the EU (which is managing the development of VB3) is, since October 2018, providing hosting and support to all Directorate Generals (DGs) of the European Commission for managing their datasets. At the time of writing the production installation of VocBench 3 at the OP is hosting 370 projects for several DGs. The number of adopters exploiting this supported hosting within the EU Commission and related institutions is rapidly growing.

FAO, which was the steward of VB2, migrated to VB3 for the maintenance of their thesaurus Agrovoc in August 2018 and has also adopted it for classification systems used in statistics. In the same domain, other institutions such as INRA, the French “Institut national de la recherche agronomique”, CIRAD, “la recherche agronomique pour le développement and, outside EU, the US National Agricultural Library (NAL) is adopting it for their NALT thesaurus and the Chinese Academy Sciences (CAS) and the Chinese Academy of Agricultural Sciences (CAAS) are exploiting VB both for their internal resources and in their collaborations with FAO.

The Senate of the Italian Republic is also adopting VocBench for the management of their thesaurus Teseo. There are then hundreds of newcomers who are already adopting the platform, starting directly with version 3, such as GelbeSeiten, the German Yellow Pages, which are using VB3 to maintain their homonymous thesaurus, and many others from all sectors.

Concerning SV, The Publications Office (OP) of the European Union is hosting a centralized installation¹⁴ of ShowVoc as a public data portal for the EU. Besides the official (and explicit) portal based on SV, the platform is also featured as an embedded service and UI component within the EU Vocabularies Portal¹⁵. Among other relevant organizations that embraced this relatively news system since its early days, we can mention FAO’s use of ShowVoc for its open data portal Caliper on FAO statistics¹⁶, followed more recently by the adoption for all of FAO

¹⁴ <https://showvoc.op.europa.eu/>

¹⁵ <https://op.europa.eu/it/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc>

¹⁶ <https://stats.fao.org/caliper/browse/showvoc/>

classifications. The LifeWatch ERIC¹⁷ consortium on BioDiversity recently introduced ShowVoc as a data hosting and publication service¹⁸, complementing its catalog – EcoPortal¹⁹ – of semantic resources on biodiversity.

6. Licensing

VocBench 3 and ShowVoc are currently distributed under the BSD 3-clause license.

LegalHTML ontology is available under the European Union Public Licence (EUPL)

7. Acknowledgements

The authors want to thank by first Johannes Keizer, once leading the “Knowledge Standards and Services” team at FAO, who initiated this VocBench endeavor (and gave the name to it!). VocBench has had different incarnations over time, but if there is a father to all of them, we know who he is. A special mention goes to Christine Laaboudi-Spoiden (currently working in Eurostat, once part of the metadata sector at the Publications Office of the EU), who first literally discovered and believed in the VocBench project, and all the people at the Publications Office and ISA2 programme (now replaced by “Digital Europe”²⁰) for their work, support and feedback on these three projects.

Sincere kuods to Carlo Marchetti, who gave the famous exception to the saying “nemo propheta in patria”, seeing in VocBench an ideal platform for the management of Teseo.

8. References

- [1] A. Stellato, S. Rajbhandari, A. Turbati, M. Fiorelli, C. Caracciolo, T. Lorenzetti, J. Keizer, and M.T. Paziienza, "VocBench: a Web Application for Collaborative Development of Multilingual Thesauri," in *The Semantic Web. Latest Advances and New Domains (Lecture Notes in Computer Science)*, F. Gandon et al., Eds.: Springer, Cham, 2015, vol. 9088, pp. 38-53, doi: [10.1007/978-3-319-18818-8_3](https://doi.org/10.1007/978-3-319-18818-8_3).
- [2] A. Stellato, M. Fiorelli, A. Turbati, T. Lorenzetti, W. van Gemert, D. Dechandon, C. Laaboudi-Spoiden, A. Gerencsér, A. Waniart, E. Costetchi, and J. Keizer, "VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons," *Semantic Web*, vol. 11, no. 5, pp. 855-881, Aug 2020, doi: [10.3233/SW-200370](https://doi.org/10.3233/SW-200370).
- [3] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, and J. Keizer, "The AGROVOC Linked Dataset," *Semantic Web Journal*, vol. 4, no. 3, pp. 341–348, 2013, doi: [10.3233/SW-130106](https://doi.org/10.3233/SW-130106).
- [4] I. Subirats-Coll, K. Kolshus, A. Turbati, A. Stellato, E. Mietzsch, D. Martini, and M. Zeng, "AGROVOC: The Linked Data Concept Hub for Food and Agriculture," *Computers and Electronics in Agriculture*, vol. 196, May 2022, doi: <https://doi.org/10.1016/j.compag.2020.105965>.
- [5] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. (2011, March) World Wide Web Consortium (W3C). [Online]. Available: <http://www.w3.org/TR/void/>
- [6] M. Fiorelli, A. Stellato, J. P. McCrae, P. Cimiano, and M. T. Paziienza, "LIME: the Metadata Module for OntoLex," in *The Semantic Web. Latest Advances and New Domains (Lecture Notes in Computer Science)*, F. Gandon et al., Eds.: Springer International Publishing, 2015, vol. 9088, pp. 321-336, doi: [10.1007/978-3-319-18818-8_20](https://doi.org/10.1007/978-3-319-18818-8_20).
- [7] World Wide Web Consortium (W3C). (2014, January) World Wide Web Consortium (W3C). [Online]. Available: <http://www.w3.org/TR/vocab-dcat/>
- [8] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy, "BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF," *Semantic Web*, vol. 4, no. 3, pp. 277-284, 2013, doi: [10.3233/SW-2012-0086](https://doi.org/10.3233/SW-2012-0086).
- [9] L. Cervone, M. Palmirani, and F. Vitali. What it is | Akoma Ntoso. [Online]. Available: http://www.akomantoso.org/?page_id=25

¹⁷ <https://www.lifewatch.eu/>

¹⁸ <https://showvoc.lifewatchitaly.eu/>

¹⁹ <http://ecoportal.lifewatchitaly.eu/>

²⁰ <https://digital-strategy.ec.europa.eu/activities/digital-programme>