JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12481

JLIS.it

The effectiveness of a Wikimedian in permanent residence: the BEIC case study

Federico Leva[(a)], **Marco Chemello**[(b)]

a) Wikimedia Italia, http://orcid.org/0000-0003-4856-0739
b) Wikimedia Italia, http://orcid.org/0000-0001-7446-8800

Contact: Federico Leva, federicoleva@tiscali.it; Marco Chemello, marco.chemello@wikimedia.it
Received: 24 April 2018; **Accepted:** 5 July 2018; **First Published:** 15 September 2018

ABSTRACT

Since 2015, Fondazione BEIC has continuously worked with a Wikimedian in residence sourced from Wikimedia Italia, showing how activity on Wikimedia projects is an integral part of an institution's work on online presence. As an update to the *Biblioteche oggi* article of 2015, where the structure and main outputs of the partnership have been described, we outline some features of an institution's work on the wikis. We'll survey some aspects of content partnerships that have been hardly described in past literature: how to plan a Wikimedian in residence; some Wikimedia Commons and Wikidata tips; other legal and technical aspects useful for a library.

KEYWORDS

BEIC; Wikipedian in residence; Wikidata; Wikimedia commons.

CITATION

Leva, F., Chemello, M. "The effectiveness of a Wikimedian in permanent residence: the BEIC case study." *JLIS.it* 9, 3 (September 2018): 141-147. DOI: 10.4403/jlis.it-12481.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12481

Since 2015, Fondazione BEIC (Biblioteca Europea di Informazione e Cultura)[1] has continuously worked with a Wikimedian in residence (WiR) sourced from Wikimedia Italia, showing how activity on Wikimedia projects is an integral part of an institution's work on online presence. As an update to the *Biblioteche oggi* article of 2015 (Consonni, Leva 2015), where the structure and main outputs of the partnership have been described, we outline some features of an institution's work on the wikis.

We'll survey some aspects of content partnerships that have been hardly described in past literature: how to plan a Wikimedian in residence; some Wikimedia Commons and Wikidata tips; other legal and technical aspects useful for a library.

Considering and planning a content partnership or Wikimedian in residence

Many institutions are interested in working with Wikimedia projects, but face difficulties in identifying what the work could comprise, or how much effort would be needed. Developing a complete plan is generally difficult or impossible without professional assistance from an experienced Wikimedia chapter or from wikimedians with a track record of such projects, but access to such assistance may be limited without a plan or budget.

In the following paragraphs we attempt to provide some insight in what an institution can plan, to facilitate a self-assessment and escape this catch 22. Having a clearer picture of the institution's purposes and strengths will allow a more fruitful interaction with Wikimedia entities, groups and individuals.

Work and coordination of WiR activities is impossible without a well-defined plan and division of labour. Depending on how much work has already been put in the institution's catalogue as opposed to the existing information in Wikimedia projects, it may be easier to start from one or the other.

Planning is especially necessary for WiRs which are slated to last only 4 months or less: such short partnerships can only be expected to produce any useful outcome if 90% of what they're going to do is defined before starting, e.g. by knowing how much and what content will be put online under a free license and what existing or missing Wikimedia content it will go towards.

Because BEIC is an effort based on an *a priori* selection of authors and works, it was rather obvious to use this selection as the foundation to work on. In simple terms, we had to know what articles and entries already existed for "our" authors, which we could then help enriching; what were missing and would need to be created from scratch; and what would probably not fit. We'd then be able to build a timeline and make sure we had steady progress towards our goals (initially for 4 months, then 6+6, then 12 etc.).

Wikidata has a key role. It is already the worldwide powerhorse and clearing house of identifiers, where all identifiers in the world are linked together. The most famous of them is probably the VIAF

[1] BEIC, https://www.beic.it/.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12481

ID, which in itself connects many national librarian identifiers. Wikidata provides numerous tools which allow handling such work efficiently, in particular *Mix'n'match* by Magnus Manske,[2] which allows to match and synchronize datasets from various institutions. Also important is the ability to mark a statement as preferred or deprecated (e.g. when alternative VIAF identifiers exist for the same person) and to survey the state of the identifier's usage with constraint violation reports (user-provided as part of the wiki-style *a posteriori* quality control, they report cases where an identifier is invalid or used for multiple entities while it shouldn't).

With Wikidata links we know exactly which and how many Wikipedia articles to work on in each language, prioritising translations and article creation as needed. Wikisource and Wikiquote are less immediate to measure, because the topics can be split differently and links go in multiple directions or are missing. The same can be said about Wiktionary and Wikibooks for common names, while Wikivoyage follows Wikipedia's structure rather closely.

Additional features of the work on Wikidata are provided *infra*.

As described later in the section on Wikimedia Commons, the BEIC releases included a bulk upload of a whole collection of images. A comprehensive release can.

In BEIC's experience, thousands of images were put into use on Wikimedia projects' articles and entries. Within 18 months from the bulk upload of the Paolo Monti photo collection, about 1500 images (9%) were used on 2500 Wikipedia articles and other Wikimedia projects, in 110 different languages, and were viewed about 4,6 million times a month.

Such a wide usage could not be imagined beforehand, nor we would have had any data or hint to select such "successful" images for upload before uploading them to Wikimedia Commons. Statistics about downloads and hotlinks[3] in Wikimedia Commons were used to surface interesting files even beyond Wikimedia.

This means that now and in the future the main way for a citizen of the world to learn about the work and legacy of this post-war notable photographer will probably go through open source, open content projects, thanks to an unstoppable, viral license.

A future development being considered is whether the entire, physical photo archive of Paolo Monti should be released under the same free license, the Creative Commons Attribution - Share Alike (CC BY-SA), even when we don't have a digitisation available. This would allow users and researchers who get a reproduction by other means (e.g. by scanning the photo themselves) to automatically have permission and be enlisted for copyleft, thereby increasing knowledge of Paolo Monti.

While european in name and cosmopolitan in scope, Fondazione BEIC has limited resources for reach out of Italy. Its multidisciplinary collections allowed to attempt serving all Wikimedia projects in all languages. Over 200 language subdomains of Wikipedia have been edited or otherwise enriched by BEIC. Not only Wikipedia, Wikidata and Commons; but also Wikisource and Wikiquote (mostly for literature), sometimes even Wikibooks and Wiktionary (mostly for illustrations of common objects).

[2] Mix'n'match, https://tools.wmflabs.org/mix-n-match.
[3] https://wikitech.wikimedia.org/wiki/Analytics/Data_Lake/Traffic/Mediacounts.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12481

The traffic to BEIC web properties is still mostly from Italian users and researchers, but with a growing interest worldwide. More importantly, Wikimedia wikis allow BEIC to serve users from hundreds of countries and cultures without investing immense sums on extensive internationalisation of its own websites.

Thanks to years of investment in training, BEIC staff is now highly independent in almost all wiki tasks, such as: categorisation, geocoding and other metadata changes; Wikidata and Wikipedia entries editing; creation and translation of Wikipedia articles; insertion of precise and informative images with captions and citation templates. It's also able to upload groups of files independently through software such as Pattypan, but for a proper result in suitable worked time there's still a need of some coding: either some librarian skills for catalog data exportation and manipulation, or some programming skills for scripts to automate the metadata and upload handling.

Eventually, the librarian and the WiR will look so integrated, and their skills so complementary, that the distinction will make little sense. In the future, every librarian will need wiki skills and work and every wikimedian will need some librarian skills. There is no reason the same shouldn't happen in other sectors, especially for any cultural institution working online (that is, all of them).

Elements of the work with Wikimedia Commons

Most content partnerships involve multimedia files, which are hosted on Wikimedia Commons. Some small tips can make a significant difference in the impact of the work.

Making thousands of images available on Wikimedia Commons is a waste of time if users cannot actually find them. The richer and more structured your original metadata is, the less work is needed in this phase.

When we uploaded in Wikimedia Commons the entire set of 17 thousands digitised photos by Paolo Monti, we had little idea what the collection really comprised: we only knew that each image was selected as representative of a set and associated with some subjects of the set; we expected that many photos of places and monuments would prove useful, especially for Wikipedia articles.

All the keywords needed to be mapped to their Commons equivalent, not least because they were in Italian rather than English. We decided to import all the keywords "as they were" and fix them directly on wiki, using powerful tools such as category redirects, SQL queries, mass editing (VisualFileChange) and autosuggest (HotCat), all available readily in the web interface of the wiki.

This work took hundreds of hours of work (being the images so many), but we received significant help from the community. Users fixed not only categories, but also metadata: for example they correctly identified unknown or mistaken places thanks to their own experience. During the 18 months following the mass upload, 235.000 edits were made to the images (an average of 14 edits per image), mostly to fix categories. While 42% of the edits were made by the BEIC team with the WiR,

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12481

52% were made by other users, and 6% by bots. More than 400 users helped, 15 of them doing more than 1000 edits.[4]

The crowdsourced work on metadata, described in the previous section, has produced significant outcomes. Many categories were created or suggested by users. We strived to have all photos categorized by:

- subject;
- place (city, region, state);
- year.

Every image was therefore tagged with 5 to 10 categories. The Paolo Monti main category on Wikimedia Commons amassed 700 subcategories, which allow a rich browsing experience and provide a nice complement to the online exhibitions (*mostre virtuali*) of the BEIC portal.

When the archive's staff and users need to find a Paolo Monti photo of a certain kind, it's now much easier for them to use the categories' galleries on Wikimedia Commons than it would be to search the catalog for relevant keywords, a process which forces them to open images individually and perform repetitive tasks hundreds of times.

Moreover, the classification on Wikimedia Commons improves by the day.

After most images had been correctly classified by depicted place, the team to try to add geographical coordinates to each of them. Using the *Locator tool*,[5] a very useful and easy-to-use geotagging interface, within 8 months we identified the original shooting place of about 50% of the Paolo Monti photographs, a very high rate, mostly with a tolerance of few meters. This "point and click" free tool uses OpenStreetMap as the cartographic base and is now effectively integrated into the Wikimedia Commons interface.

Geotagging a photo collection, and in particular an historical collection, is very useful, not only for research purposes, but also for presenting it to a larger audience. BEIC created a prototypal interface, called "Paolo Monti - His legacy",[6] to let the public navigate through the collection on a geographical map, where each layer is generated by a different category on Wikimedia Commons. So the map is changing and refining every day, with the help of other users.

We think that allowing a collection of historical photos to be browsed in such a rich means giving it a new life.

[4] https://commons.wikimedia.org/wiki/File:Dare_spazio_-_condividere_e_georeferenziare_gli_archivi_fotografici_-_Il_caso_Paolo_Monti.pdf.
[5] https://commons.wikimedia.org/wiki/Commons:Locator-tool.
[6] http://www.beic.it/mappa-paolo-monti/.

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12481

Benefits of Wikidata

Thanks to Wikidata, not only we could store identifiers we already knew about for our content (such as VIAF and CERL) but we also found out more (like DBI, *Dizionario Biografico degli Italiani* by Treccani). Identifier coverage was improved with the help of easy wiki technologies and of wiki volunteers.

Such cross-linking will allow in the future to provide a real-time and free integration of links to select authority controls, as well as discovery tools advancements such as alternative names for our authors (like Italian or Russian names for authors whose name is canonically written in Latin), simply by fetching the most up to date Wikidata data connected to our authors, and eventually their works.

An ultimate goal for a digital library like BEIC is, in fact, to migrate data of its whole collection of books into Wikidata, taking advantage of his LOD (Linked Open Data) environment. Once the metadata about authors has been transferred, we can work on migrating also publishers' data, then we will be able to migrate all bibliographical references, linking them with frontispieces and book illustrations already uploaded into Wikimedia Commons. Wikipedia needs a lot of reliable references, but Wikimedia still lacks of a vast bibliographical database, so we can help creating it.

Some unexpected benefits of the wiki way

One of the defining features of MediaWiki is that pages are tightly connected by a number of links, which take various forms and are highly tracked. Categories are an easy way to establish links between a cluster of items such as a group of related images, but such a cluster will remain lonely and potentially unused if it's not connected to the rest of the wiki. Hence the categories, if created on purpose, need to also be part of the existing category tree: for instance there's little use for a category "Photos by Paolo Monti in northern Italy", while it's useful to have "Photos by Paolo Monti in Milan" which is then a subcategory of the existing categories for Milan and the higher administrative levels above it.

Similarly, when creating new articles on Wikipedia, an important and often overlooked need is that they should link other articles (be "wikified") but also be linked by other articles, otherwise they're considered "orphan" by Wikipedian standards. Originally, an orphan article was thought to be almost impossible to find, since web crawlers would not find web pages without incoming links and few people used the internal search engine of the wiki. Nowadays, orphan pages can still be found on main search engines, which have special indexers for Wikipedia, but they're still in a rather weak position. First, an article without incoming links is not "wanted" by other articles and can be felt as less important; second, an article which is not interconnected with others can be felt as an extraneous element, which doesn't fit into the encyclopedia, and therefore its chances to be deleted are higher.

Creating links between articles is a wiki task, but it's just a variant of a task which suits librarians: finding similar topics or related authors and connecting them. When writing a biography about an 18th century mathematician, looking for an incoming link means for instance to:

JLIS.it 9, 3 (September 2018)
ISSN: 2038-1026 online

DOI: 10.4403/jlis.it-12481

- search authority records like CERL for name variants and create redirects from them, so that any incoming link to those variants reaches its intended destination;
- find what other authors or persons have inspired, are mentioned or are otherwise connected to the person's work, mention the fact in both articles to link them;
- find related events, inventions or concepts and explain the person's contribution or connection to them, with a reciprocal link emerging.

This kind of work helps focus the article on information which are of interest for the wiki, because they were already there, and which help users find more of what interests them among our content, even if they had no idea they were looking for it. Without connections, there's a risk of developing content which only caters to the existing audience of the institutions or is nevertheless *autoreferenziale* or "preaches to the converted".

Thanks to Wikimedia users, as we knew would happen thanks to the Bundesarchiv experience, a number of corrections have been made to the catalog data, without an expensive reverification of every entry.

Interesting corrections of the original metadata include:

- fixing the rare misattribution of authorship for a work attributed to a similarly named person in the same period (e.g. son or brother);
- fixing the subject or the shooting place of an historical photograph by Paolo Monti.

Increasing demand for usage of BEIC content by third parties has driven an overhaul of the copyright status metadata and of the copyright legal notes. Such work has proven useful also for integration into Europeana. Following the public domain charter of Europeana, itself inspired by the public domain manifesto, a general rule was established that data and public domain works remain in the public domain (CC0), that what Fondazione BEIC owns some copyright on is distributed under copyleft (CC BY-SA) and that the rest is marked as copyrighted (with rightsstatements.org URIs if possible). This spares time of the BEIC staff and helps spreading contents worldwide.

References

Consonni, Chiara, and Federico Leva. 2015. "Progetto GLAM/BEIC." *Biblioteche Oggi*, vol. 33 (marzo): 47–50.