May 24, 2021

Protocol of the benchmark for designing and developing the National Edition of Aldo Moro's works

Sebastian Barzaghi[1]

[1]University of Bologna

Sebastian Barzaghi: Digital Humanitites Advanced Research Centre (/DH.arc)

In Development Share dx.doi.org/10.17504/protocols.io.bupxnvpn

Sebastian Barzaghi

ABSTRACT

This work aims to define a series of reference models that would serve as a guide to the design and development of the National Edition of Aldo Moro's works. These models have been defined through a benchmarking process that is organized as follows:

- **content analysis** on a sample of 30 digital editions, evaluated on the basis of certain criteria defined in [Sahle 2014];
- **processing** of the data gathered as a result of the content analysis, so as to extract the relevant information, and visualize it;
- **review** of the data processing results and consideration of the models that can be used as reference.

The digital editions that meet the quality criteria taken into consideration are equipped with the following characteristics: their audience is composed of both domain experts and generic users; their documentation is rich and accurate; their content is described by a complete set of metadata; they and their single parts are citable and uniquely identifiable; their data model is geared towards interoperability and interlinking between its contents and the relevant resources already existing on the Web; they use visualization and storytelling tools so as to convey information intuitively; their information architecture is well-structured and easily navigated; their data and contents can be downloaded in many different formats; they take advantage of Open Source software and tools; finally, their contents are open and accessible to anyone.

DOI

dx.doi.org/10.17504/protocols.io.bupxnvpn

PROTOCOL CITATION

Sebastian Barzaghi 2021. Protocol of the benchmark for designing and developing the National Edition of Aldo Moro's works. **protocols.io** https://dx.doi.org/10.17504/protocols.io.bupxnvpn

KEYWORDS

digital editions, python, data visualization, benchmark, content analysis, digital humanities, digital scholarly editions

CREATED

May 01, 2021

LAST MODIFIED

May 24, 2021

PROTOCOL INTEGER ID

49623

MATERIALS TEXT

The materials used in the protocol are:

- Patrick Sahle; in collaboration with Georg Vogeler and the members of the IDE; Version 1.1, June 2014 (Version 1.0, September 2012 – January 2014; German version 1.1:http://www.i-d-e.de/publikationen/weitereschriften/kriterien-version-1-1/);
- Barzaghi, Sebastian. (2021). Benchmark Dataset for designing and developing the National Edition of Aldo Moro's works [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4779123
- Barzaghi, Sebastian. (2021, May 21). Benchmark for designing and developing the National Edition of Aldo Moro's works. Zenodo. http://doi.org/10.5281/zenodo.4779139

BEFORE STARTING

This protocol takes for granted some basic knowledge about what a digital edition is, how to program with Python, and how to use Jupyter Notebooks. Before you start, remember to check whether Python3+ is installed on your computer or not. If it is not yet installed, please check Python official installation guidelines and make sure to follow them accordingly. You will also have to install the following libraries on your machine in order to correctly execute the Python code snippets that are present in the Jupyter Notebook:

- numpy
- pandas
- matplotlib
- seaborn

Data collection

1 A sample of 30 editions has been selected for their analysis and evaluation on the basis of certain criteria. The evaluation criteria that have been used to review the sample are based on part of the criteria for reviewing scholarly digital editions compiled by Patrick Sahle in collaboration with Georg Voegler and IDE (Institut für Dokumentologie und Editorik) members (Sahle 2014).

1.1 The aspects that have been taken into consideration during the evaluation process are the following:

- **Documentation** (*Documentation*, *Scholarly objectives*, *Mission* focusing on the objectives, *Documentation and associated texts*);
- **Audience** (*Mission*, focusing on the audience);
- **Representation** (*Representation of documents and texts*);
- **Data model** (*Data modelling*);
- **Browse**;
- **Search**;
- **Indices**;
- **Quality of the presentation**;
- **Metadata** (*Metadata for description of and interlinkage between objects in the edition*);
- **Identification** (*Identification and citation*);
- **Technical interfaces**;
- **Formats** (*Spin offs and export formats*);
- **OS-OA** (*Access to basic data*, *Rights and licences*);
- **Additional features**;

1.2 Each aspect (except for *Audience* and *Data model*) has been given a score between **0**, **0.5**, and **1**, where:

- **0** represents a value that witnesses either the total absence or the lack of quality of the edition in terms of that specific aspect;
- **0.5** represents a value that witnesses a suboptimal implementation of that specific aspect in the edition;
- **1** represents a value that witnesses an optimal implementation of that specific aspect in the edition.

Citation: Sebastian Barzaghi (05/24/2021). Protocol of the benchmark for designing and developing the National Edition of Aldo Moro's works. https://dx.doi.org/10.17504/protocols.io.bupxnvpn

1.3 The values associated with each aspect of each edition, along with its title and URL, have been collected in a CSV dataset (**benchmark-moro.csv**), in which each row represents a digital edition and each column represents one of its aspects.

benchmark-moro.csv

1.4 The CSV dataset is structured as follows:

- **name**: the title of the digital edition;
- **url**: the URL address of the digital edition;
- **documentation**: the score assigned to the documentation of the edition. If the score is **0**, the documentation is non-existent or insufficient; if the score is **0.5**, the documentation is partial and contains superficial information about the project, its contents, and its technical implementation; if the score is **1**, the documentation is complete and contains detailed information about the project, its contents, and its technical implementation;
- **audience**: the type of audience of the edition. The value **ns** indicates the audience is not specified; the value **specialist** indicates domain experts, such as researchers, scholars, etc.; the value **general** indicates laymen; the value **both** indicates both domain experts and laymen;
- **text-representation**: the score assigned to the textual representation in the edition. If the score is **0**, the textual representation is low-quality; if the score is **0.5**, the textual representation is medium-quality; if the score is **1**, the textual representation is high-quality;
- **data-model**: the type of data model used in the edition. The value **TEI** indicates that the digital edition's data model follows the Text Encoding Initiative (TEI) Guidelines; the value **RDF** indicates that the digital edition's data model is based on the Resource Description Framework (RDF); the value **other** indicates that the digital edition's data model is another one;
- **browse**: the score assigned to the browsing functionalities of the edition. If the score is **0**, browsing is difficult and does not allow rapid access to contents; if the score is **0.5**, browsing is fairly difficult and allows rapid access to a limited amount of content; if the score is **1**, browsing is easy and allows rapid access to contents;
- **search**: the score assigned to the search functionalities of the edition. If the score is **0**, the search system is non-existent, or does not allow query refinement; if the score is **0.5**, the search system allows limited query refinement; if the score is **1**, the search system allows complex query refinement;
- **indices**: the score assigned to the use of indexes by the edition. If the score is **0**, the digital edition has no indexes; if the score is **0.5**, the digital edition has a few indexes, traditionally implemented; if the score is **1**, the digital edition has an adequate number of indexes that represent an alternative way to navigate the site;
- **presentation-quality**: the score assigned to the quality of presentation of the edition's contents. If the score is **0**, the digital edition presents its contents as flat text, taking no advantage of digital technologies to visualize additional features; if the score is **0.5**, the digital edition presents its contents as text that shows a few additional features by taking advantage of digital technologies; if the score is **1**, the digital edition presents its contents as multidimensional text that shows additional features by taking full advantage of digital technologies;
- **metadata**: the score assigned to the quality of the edition's metadata. If the score is **0**, the contents of the digital edition are not accompanied by any significant metadata or just by a few of them; if the score is **0.5**, the contents of the digital edition are accompanied by a limited number of significant metadata; if the score is **1**, the contents of the digital edition are accompanied by a fair amount of significant metadata;
- **identification**: the score assigned to the edition's mechanisms for its identification and citation. If the score is **0**, the digital edition has no permanent identifiers at any level of granularity and does not provide any clear indication for its citation model; if the score is **0.5**, the digital edition may have permanent identifiers at the highest levels of granularity and may provide a suggestion for its

Citation: Sebastian Barzaghi (05/24/2021). Protocol of the benchmark for designing and developing the National Edition of Aldo Moro's works. https://dx.doi.org/10.17504/protocols.io.bupxnvpn

citation model; if the score is **1**, the digital edition has permanent identifiers at every level of granularity and provides clear indications for its citation model;

- **tech-interfaces**: the score assigned to the edition's technical interfaces. If the score is **0**, the digital edition does not use any technical interface to allow data reuse; if the score is **0.5**, the digital edition uses an undocumented technical interface to allow data reuse; if the score is **1**, the digital edition uses a documented technical interface to allow data reuse;
- **formats**: the score assigned to the availability of different download formats for the edition's contents. If the score is **0**, the digital edition do not support content download in any format; if the score is **0.5**, the digital edition supports at least one format for downloading content, such as PDF or eBook; if the score is **1**, the digital edition supports multiple formats for downloading content, including structured formats like XML, JSON, ecc.;
- **os-oa**: the score assigned to the terms of use and licences to access and reuse the contents and other assets of the edition. If the score is **0**, the software used to develop the digital edition is not Open Source, the edition's contents cannot be reused by other researchers, and, in order to access them, the users have to sign in and sometimes pay a subscription fee; if the score is **0.5**, the software used to develop the digital edition can be Open Source and the edition's contents are freely accessible, but not reusable; if the score is **1**, the software used to develop the digital edition is Open Source, and the edition's contents are freely accessible and reusable;
- **add-features**: the score assigned to the additional features of the edition. If the score is **0**, the digital edition does not provide additional features for displaying its contents; if the score is **0.5**, the digital edition provides basic additional features for displaying its contents; if the score is **1**, the digital edition provides complex additional features for displaying its contents.
- **total**: the total score assigned to the edition.

Data analysis

2 A Jupyter Notebook has been created to document the whole benchmark process, including the code snippets used to produce visualizations.

Benchmark for designing and developing the National Edition of A 1

source by Sebastian Barzaghi

2.1 The libraries that have been imported to conduct the analysis are:

- **numpy** and **pandas** for data processing;
- **matplotlib** and **seaborn** for data visualization.

2.2 For each aspect, the examined editions have been grouped and counted according to the score or value that was assigned to them in the CSV dataset; then, the results have been visualized in a barplot with the following characteristics:

- a descriptive **title** summarizing the results;
- a **bottom spine**, listing the three possible values (0, 0.5, and 1) and a **label** indicating the considered aspect;
- three **bars**, each representing the amount of editions belonging to that specific score and colored accordingly (0=red, 0.5=yellow, 1=green, literal value=purple), with the amount of editions shown on the top;

Results review

3 For each edition, its total score has been plotted in a **horizontal bar chart** to better visualize the final results in a descending order.

Citation: Sebastian Barzaghi (05/24/2021). Protocol of the benchmark for designing and developing the National Edition of Aldo Moro's works. https://dx.doi.org/10.17504/protocols.io.bupxnvpn

3.1 On the basis of the final results, a subset of editions has been selected to serve as reference models for design and development. In the case of the National Edition of Aldo Moro's works, the selected editions are those that received an optimal score (greater than or equal to 8).

4 In order to better understand their strengths and weaknesses, a series of **radar graphs** have been created for the selected editions.

Publication

5 The CSV dataset and the Jupyter Notebook have been published on Zenodo and released with a Creative Commons Attribution 4.0 International license.

Citation: Sebastian Barzaghi (05/24/2021). Protocol of the benchmark for designing and developing the National Edition of Aldo Moro's works. https://dx.doi.org/10.17504/protocols.io.bupxnvpn