IFLA WLIC 2022 DUBLIN

Submitted on: 14.06.2022

Cataloger acceptance and use of semiautomated subject recommendations for web scale linked data systems

Jim Hahn
University Libraries, University of Pennsylvania, Philadelphia, USA.
E-mail address: jimhahn@upenn.edu

Abstract:
As catalogers begin to integrate linked data descriptions into large-scale discovery graphs through RDF editors, interventions such as semi-automated subject description (http://lcsh.annif.info) are extending and supporting their professional expertise. A large corpus of 9.3 million (9,304,455) title and subject pairs from the IvyPlus Platform for Open Data (POD), along with SVDE bibliographic data, were used for training a semi-automated subject indexing tool for use in BIBFRAME linked data editors. Thereafter, catalogers evaluated the automated subject outputs for inclusion in their descriptions of BIBFRAME resources and the general usefulness of semi-automated subject suggestions. This paper presents the findings of a mixed-methods inquiry to better understand catalogers' preferences for incorporating machine learning outputs into their work.

Keywords: Linked Data, Machine Learning, Big Data, Metadata, Ethics

Introduction

BIBFRAME implementation poses challenges related to resources, scale, and training. The Share-VDE (SVDE) system is an implementation of a large-scale linked data BIBFRAME network. SVDE is a library-driven initiative that brings together the bibliographic data and authority files of a community of libraries in a discovery environment, based on linked data in general and using BIBFRAME description specifically. Included in the SVDE System are linked data descriptions from University of Pennsylvania, Stanford University, Library of Congress, National Library of Norway, and the Smithsonian, which includes 42.3 million records processed (29,580,361 million bib. + 12,757,192 authority records). The BIBFRAME data can be searched at https://svde.org. Given the large network of nodes with which to place new descriptions of BIBFRAME entities, semi-automated methods have been proposed (Hahn, 2021) to support and extend professional catalog expertise.

Professional library catalogers evaluated the automated subject outputs for inclusion into their descriptions of BIBFRAME Work resources and evaluated the usefulness of semi-automated subject suggestions in general. This paper will detail cataloger preferences for semi-automated cataloging using Annif software, loaded with the LCSH vocabulary (http://lcsh.annif.info) as a base for subject assignment support within linked data editors, such as the Sinopia RDF editor. Considerations for ethical use of both subject vocabularies and for aspects of hybrid "human in the loop" automation in

cataloging are presented. Technology-intensive fields have evaluated training in professional ethics (Fiesler, Garrett, and Beard, 2020), and this paper will foreground ethical considerations for semi-automated subject description.

The LCSH vocabulary is not a neutral dataset (Noble, 2018), and this exploration is not neutral in its goals, which are to support cataloging professionals. Foregrounding the consideration of AI's ethical needs is an idea most recently articulated by Birhane, who stated that "Ethical practice, especially with regard to algorithmic predictions of social outcomes, requires a fundamental rethinking of justice, fairness, and ethics above and beyond technical solutions." (2021, p. 8) There is evidence to suggest that for many users, classification language has and continues to use problematic language that yields non-inclusive results; Noble (2018) noted that assigning a subject to a person or group can result in further objectification of groups of people, particularly marginalized or minoritized groups. While the goals to semi-automated subject assignment have been the subject of ample experimentation, a parallel but inverse process could, in the future, begin to systematically autosuggest both inclusive language and subject vocabularies based upon heretofore excluded or overlooked datasets. Such data incorporation will require further development, likely using frames of relational ethics to incorporate such a feature.

The ethics of using a problematic vocabulary pose implementation concerns. In *Data Feminism*, D'Ignazio and Klein considered this context through the lens of data feminism, which "… asserts that data are not neutral or objective. They are products of unequal social relations, and this context is essential for conducting accurate, ethical analysis." (D'Ignazio and Klein, 2020, p. 18) The contextual framework is essential to our understanding of the ethics of semi-automated subject description, the implementation of which is poised for production; contextual considerations for this principle of data feminism include "… a process that includes understanding the provenance and environment from which the data was collected, as well as working hard to frame context in data communication." (D'Ignazio and Klein, 2020, pp. 171–172) Fitzpatrick asked academics to approach discourse with a listening lens and to consider negative effects while assuming positive intent (Fitzpatrick, 2019). Using generous thinking, this paper articulates generous machine learning operations in libraries. In this paper, the idea of incorporating the expertise of a "cataloger in the loop" is indicative of the type of approach that may serve to support the cataloger's needs while using automation in ways that can truly help professionals complete resource-intensive but ultimately valuable work. The inquiry herein of semi-automated use by catalogers includes understanding the range of human needs in RDF creation with the BIBFRAME vocabulary. Having established ethical grounding to the project of semi-automated support of cataloger needs in linked data description, the paper next explores the relevant literature that contextualized our inquiry into cataloger preferences and professional acceptance of novel tools, processes, and practices.

Background

Bennett et al. (2014) described subject assignment as "a combination of intellectual and manual tasks." (p. 42) Challenges to assigning subjects for resources are not a novel problem as this issue has been a focus for sustained research, beginning with the advent of online public access catalogs; for instance, in a 1986 article on the nature of subject access, Bates underscored that "… it is practically impossible to instruct indexers or catalogers how to find subjects when they examine documents." (p. 360) In the same article, Bates also reflected that the "… Newtonian/mechanistic assumption has been that somewhere there is an ideal indexing system or language that will enable us to produce the one perfect description or set of descriptions for each document. These ideal descriptions will, in turn, produce the best possible match with users' needs as expressed in queries. Each improvement in human or machine indexing is to take us closer to that ideal. But suppose instead that that ideal is impossible in principle, because both indexing behavior and information searching behavior are at least in part indeterminate and probabilistic." (p. 360) Bates presented an uncertainty principle in subject indexing. The difficulties in assigning subjects to resources continued to be a focus of "expert systems" iteration.

Hawks reported on the "OCLC Automated Title Page Project," describing it as "One of the most significant projects" at the time (1994) in the arena of expert systems. According to Hawks, "OCLC's study examines the viability of scanning title page information into an automated cataloging system. The system would evaluate this data and produce a first-level bibliographic description as defined by AACR2." (p. 205) Other early exemplars included a "cataloging adviser": "MAPPER was developed for two purposes: to make expert advice available to novice map catalogers and to improve conventional instruction in map cataloging." (p. 206) Contemporary automated subject indexing methods were explained by Golub (2021) as "… the application of supervised machine learning algorithms. In this approach the algorithm 'learns' about characteristics of target index terms based on characteristics of documents (e.g., frequently occurring tokens) that have already been manually indexed with those terms; these documents are called training documents." (p. 706) Suominen (2019) wrote on the development of Annif software package used for automated subject indexing, "we have developed Annif, an open-source multi-algorithm auto-mated indexing tool. After loading a subject vocabulary and existing meta-data, Annif learns how to assign subject headings to new documents. It can also be used as a web service that can be integrated with other systems." (p. 2) The Annif software package facilitated the present study of autosuggesting LCSH subject headings.

Mixed-methods inquiry

An informal mixed-methods approach was used to ascertain cataloger preferences in utilizing the Annif LCSH subject suggestions. Methods included a survey of professional catalogers at the University of Pennsylvania Library, feedback from the LD4 Sinopia user group, and conference discussion at the Semantic Web in Libraries Conference, SWIB.

Internal surveys at University of Pennsylvania
A survey of catalogers who have been evaluating linked data editors was distributed in early 2022. These catalogers receive monthly training in linked data focusing on the BIBFRAME vocabulary. The survey is available in Appendix 1. Thematic areas of inquiry for the survey included ascertaining the importance of relevance, website response time, targeting subjects based on genre of resource, and asking catalogers if they would use such a resource in their work. Finally, as an open-ended response, the survey collected ideas for features, which might be incorporated into future versions of the Annif subject suggestion service.

Sinopia User Group discussions: LD4 Affinity Group discussions
During the April 2021 meeting of the Sinopia User Group, the project of Annif subject suggestions was introduced, and feedback was requested for development. The presentation included an overview of how data might be used in the Sinopia RDF editor in the future, in response to feature requests aimed at streamlining descriptions in the RDF editor. The cohort of the Sinopia User Group is comprised of professionals involved in BIBFRAME vocabulary training or RDF editing. The group is generally comprised of professionals seeking to understand features of functionality of the RDF editor.

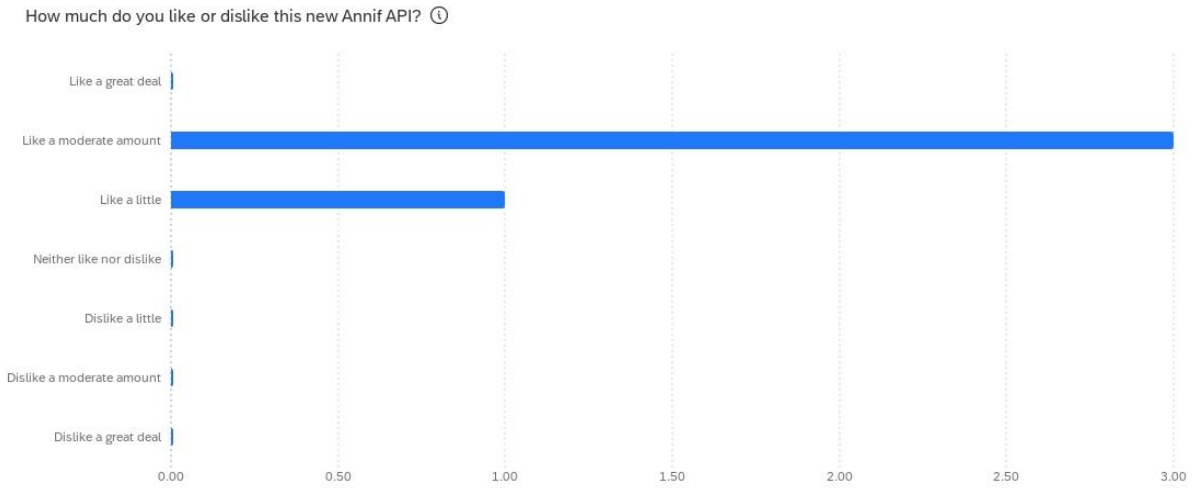
Conference discussions: Semantic Web in Libraries SWIB 2021
The conference presentation on Annif using LCSH vocabulary, held at Semantic Web in Libraries 2021, included a technical discussion on training best practices with Annif. The audience of the Semantic Web in Libraries fosters international participation and includes linked data professionals from Europe and further afield, as well as technical staff and generalists with an interest in linked data for application in libraries.

Inquiry results

Internal surveys at the University of Pennsylvania Libraries
Results from the surveys of linked data catalogers at the University of Pennsylvania Libraries showed a desire to develop subject recommendations that could target a particular language. The initial implementation did not specifically differentiate among languages; in practice, the training data used

all text and associated subject assignments. In particular, the catalogers made features for subject suggestions based on input language.

Response Table Q1. How much do you like or dislike this new Annif API?



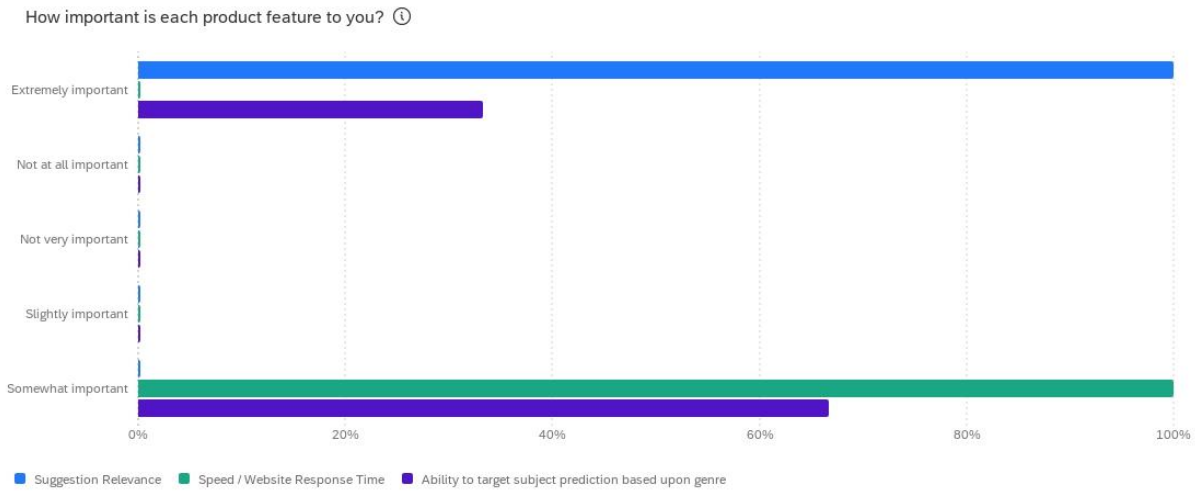
Responses to Q2. What do you like most about this new Annif API?

- easy to use, fast response
- returns some useful suggestions for subject headings
- clean design
- ease of use

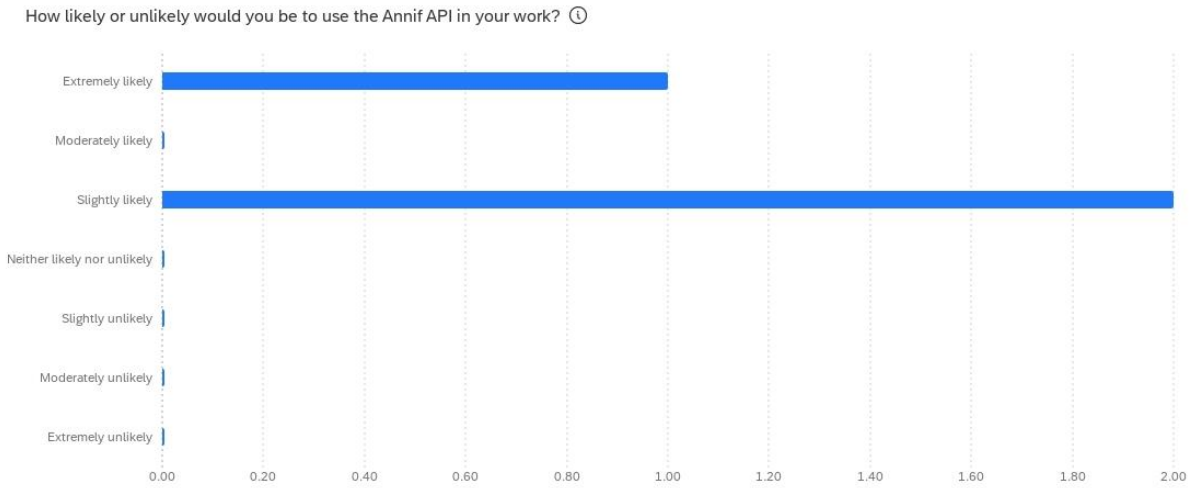
Responses to Q3. What do you like least about this Annif API?

- returns some irrelevant subject headings
- would be useful to be able to specify the language of the text that one enters

Response Table to Q4. How important is each product feature to you?



Response Table Q5. How likely or unlikely would you be to use the Annif API in your work?



What features would you like to see in the next version of the Annif API service?

- specify text language
- functionality to pick language and limit to geographic area

Sinopia User Group discussions: LD4 Community Group
The Sinopia User Group discussion included a feature request to add support for MeSH-RDF, the Medical Subject Heading from National Library of Medicine. A discussion on how foreign language titles might be handled using new or updated data training was held, and an idea for improving the training process by using additional information beyond the title, such as summaries, was presented.

Conference discussions: Semantic Web in Libraries/SWIB 2021
The conference presentation at Semantic Web in Libraries 2021 on Annif using LCSH vocabulary included a technical discussion on training best practices with Annif. The theme of the discussion was predominantly technical, but it also focused on the various ways to improve training data. One improvement suggested by the SWIB audience included working on evaluating the use of both 245a and 245b (title and subtitle, where available) of source training data. The initial data training focused on the 245a title and excluded other text from the title. There was also discussion on how to manage the ongoing training needs of the system where the LCSH vocabulary is updated. Thematically, the process of integrating ongoing data training falls into the domain of the emergent field of machine learning operations.

Discussion

Schultz-Jones et al.'s (2012) study suggested that with the advent of next-generation catalog arose a corresponding shift in perceptions of cataloging quality among catalogers: "the shift in the cataloguer's judgment from rigid standards for transcription to meeting the requirement for more metadata that matches the user need of find-ability." (p. 49) This aligns with the Annif feedback that was collected, while informal and representative aspects of convenience samples contained feedback on aspects of quality in general and on the API functionality to generate accurate subject recommendations based on input text. The general-purpose Annif API for subject suggestions in LCSH can be improved by way of selecting and, in some cases, further curating appropriate training data to align with the requested resource. Therefore, ideas for training improvement discussed herein include reworking the general API from lcsh.annif.info into bifurcated specialized language and genre-specific APIs.

Multilingual subject recommendations

Regarding language feature requests, Riva (2022) considered topical subjects and classification within multilingual catalogs, noting that "classification is enticing as a language switching hub, because the classification notations may appear to be language-neutral, but there are cultural expectations built-in to the design of classification, as basic as what topics go together, and which do not." (p. 95) The process of undertaking language associations in subjects does require human-curated mapping to ensure success; as Riva stated, "subject heading languages and thesauri also need to grapple with the issue that what is or is not viewed as being the same topic differs between language or cultural groups, even when the formal structures of the schemes are compatible. Linking pre-existing subject schemes devised according to different structures may best be described as a mapping process. When subject heading mappings have been carefully curated by bilingual cataloguers and the subject heading languages are compatible in structure, the results can be very good." (p. 95) Finding compatible subject structures for training and appropriate levels of curation will be important for refining the Annif API for language-based suggestions. Riva also mentioned the European MACS (multilingual access to subjects) project as another exemplary success in this area (Riva, 2022; Landry, 2004). This may be a starting point for incorporating an Annif LCSH API with that is "language aware."

Genre-based recommendations

Lee and Zhang addressed the difficult concept of genre in cataloging by underscoring that "modern genre theorists reject the singular notion of genre as rigid text categories and maintain that genres are forms of social action." (2013, p. 892) Overall, the data for genre may not be systematically unified; as Lee and Zhang's study, *Tracing the Conceptions and Treatment of Genre in Anglo-American Cataloging* (2013), reported in detail, "The study traced the conceptions and treatment of genre in modern Anglo-American cataloging through an analysis of four sets of rules: Panizzi's 91 Rules, Cutter's Rules for a Printed Dictionary Catalog, AACR2R, and RDA. Genre appears in all four, ranging from being a vague and minor concern in the 91 Rules to becoming an important attribute of the work entity and a useful indexing element in RDA. The cataloging encoding authorities have preferred 'form' to 'genre,' but failed to provide a rigorous and useful definition for either or a clear and consistent distinction between the two." (Lee & Zhang, 2013, p. 909) As a systematic and formal approach to genre is not found in the training data, the present data need human curation to determine whether they can be usefully applied. Therefore, genre as a targeted Annif API may be explored on a catalog-by-catalog basis and by way of human-curated corpus construction. If the Annif genre API is to be developed, and become useful and used by catalogers, it will require ongoing feedback and analysis to understand if genre can be trained in the Annif system.

Algorithm improvements

The Annif software has been continually developed and improved. There are now baseline reports of algorithm ensembles and their respective rankings that were not previously available (Suominen et al., 2022). Reports of the most useful combination of algorithms can be valuable for improving the machine learning aspects of the Annif API redevelopment. Specifically, the original TF-IDF algorithm can be replaced with an ensemble of two algorithms that were reported in Suominen et al. to have higher scores for subject recommendations for the purpose of subject assignment to bibliographic works (2022). The next combination of algorithms that will be used together in the Annif LCSH API includes the fastText algorithm (Joulin et al., 2016) and the Omikuji Bonsai tree-based machine learning algorithm (Khandagale et al., 2020), both of which are packaged in the latest release of Annif software.

Conclusion

The ways in which Annif may be valuable to catalogers will partially depend on the ability for Annif functions to generate appropriate data for the task at hand. A general-purpose Annif holds limitations toward accuracy, whereas bifurcating into several language and perhaps genre-specific Annif APIs would be better poised to target the cataloger's context of the resource, and the RDF editor may dynamically choose that API, which is closer to the actual need of the cataloger based on that inferred

context. This will hopefully contribute to alleviating some of the issues associated with higher precision of machine learning based recommendations, while at the same time allowing catalogers to best utilize the expertise in professional cataloging.

In this paper, the idea of incorporating the expertise of a "cataloger in the loop" is indicative of the type of approach that may serve to support the cataloger's needs, while using automation in ways that can truly help professionals complete time-consuming but ultimately valuable work. The semi-automated process was inspired by listening to catalogers' needs in RDF creation. It is hoped that the research herein responded to observed difficulties and provided discourse with those early adopters who were experimenting with novel linked data RDF editors.

References

Bates, M. (1986). Subject access in online catalogs: A design model. *Journal of the American Society for Information Science*, *37*(6), 357.
DOI: https://doi.org/10.1002/(SICI)1097-4571(198611)37:6<357::AID-ASI1>3.0.CO;2-H.

Bennett, R., O'Neill, E. T., and Kammerer, K. (2014). AssignFAST: An autosuggest-based tool for FAST subject assignment. *Information Technology & Libraries*, *33*(1), 34–43.
DOI: https://doi.org/10.6017/ital.v33i1.5378.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 8.
DOI: https://doi.org/10.1016/j.patter.2021.100205.

D'Ignazio, C., and Klein, L.F. (2020). *Data Feminism*. MIT Press.

Fiesler, C., Garrett, N., and Beard, N. (2020). What do we teach when we teach tech ethics: A syllabi analysis. *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, Portland, OR, 289–95.
https://doi.org/10.1145/3328778.3366825.

Fitzpatrick, K. (2019). *Generous Thinking: A Radical Approach to Saving the University*. John Hopkins University Press.

Golub, K. (2021). Automated subject indexing: An overview. *Cataloging & Classification Quarterly*, *59*(8), 702–719.
DOI: 10.1080/01639374.2021.2012311.

Hahn, J. (2021). Semi-automated methods for BIBFRAME Work entity description, *Cataloging & Classification Quarterly*, *59*(8), 853-867.
DOI: 10.1080/01639374.2021.2014011.

Hawks, C. P. (1994). Expert systems in technical services and collection management. *Information Technology and Libraries*, *13*(3), 203–212.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification, *arXiv: 1607.01759*, 3.
https://arxiv.org/abs/1607.01759.

Khandagale, S., Xiaom, H., and Babbar, R. (2020). Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning*, *109*, 2099–2119.
DOI: https://doi.org/10.1007/s10994-020-05888-2.

Landry, P. (2004). Multilingual subject access: The linking approach of MACS, *Cataloging & Classification Quarterly*, *37*(3–4), 177–191.

DOI: 10.1300/J104v37n03_11.

Lee, H.-L., and Zhang, L. (2013). Tracing the conceptions and treatment of genre in Anglo-American cataloging. *Cataloging & Classification Quarterly*, *51*(8), 891–912.
DOI: 10.1080/01639374.2013.832457.

Noble, S. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 134–152.

Riva, P. (2022). The multilingual challenge in bibliographic description and access. *JLIS.it*, *13*(1), 86–98.
DOI: https://doi.org/10.4403/jlis.it-12737.

Schultz-Jones, B., Snow, K., Miksa, S., and Hasenyager, R.L., Jr. (2012). Historical and current implications of cataloguing quality for next-generation catalogues. *Library Trends*, *61*(1), 49–82.
DOI: https://doi.org/10.1353/lib.2012.0028.

Suominen, O. (2019). Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, *29*(1), 1–25.
DOI: https://doi.org/10.18352/lq.10285.

Suominen, O., Inkinen, J., and Lehtinen, M. (2022). Annif and Finto AI: Developing and implementing automated subject indexing. *JLIS.It*, *13*(1), 265–82.
DOI: https://doi.org/10.4403/jlis.it-12740.

Appendix 1

Q1 How much do you like or dislike this new Annif API?

- ○ Like a great deal (1)
- ○ Like a moderate amount (2)
- ○ Like a little (3)
- ○ Neither like nor dislike (4)
- ○ Dislike a little (5)
- ○ Dislike a moderate amount (6)
- ○ Dislike a great deal (7)

Q2 What do you like most about this new Annif API?

Q3 What do you like least about this Annif API?

Q4 How important is each product feature to you?

	Extremely important (1)	Somewhat important (2)	Slightly important (3)	Not very important (4)	Not at all important (5)
Suggestion relevance (1)	○	○	○	○	○
Speed/website response time (2)	○	○	○	○	○
Ability to target subject prediction based upon genre (3)	○	○	○	○	○

Q5 How likely or unlikely would you be to use the Annif API in your work?

- ○ Extremely likely (1)
- ○ Moderately likely (2)
- ○ Slightly likely (3)
- ○ Neither likely nor unlikely (4)
- ○ Slightly unlikely (5)
- ○ Moderately unlikely (6)
- ○ Extremely unlikely (7)

Q6 What features would you like to see in the next version of the Annif API service?