FEATURE

Bifurcation of Semi-Automated Subject Indexing Services

Jim Hahn

Semi-automated subject indexing methods use attributes from metadata descriptions as training data. A survey to shape inclusion of metadata attributes that align a machine learning model within a contextual linguistic domain generated the initial genre targets for experimentation. The second part of this study then tested the genre attributes from the survey. These bifurcations (or branching points) served as the basis for machine learning model development and evaluation. The machine learning models in semi-automated indexing systems are the drivers of the automated subject outputs. The initial results of this multipart experiment indicate that measures of the mean precision and recall (the F1 metric) improved for several—but not all—types of genres that were of interest to knowledge workers.

Knowledge workers face an ever-increasing expansion of the knowledge universe. Therefore, skilled professionals require increased support to extend their specialized expertise, and semi-automated tools based upon machine learning may help support their subject-description tasks. Semi-automated subject indexing offers opportunities to reference and extend professional cataloger skills—in contrast to completely automated support—as complete automation has eroded some professionals' skills.[1] Golub reported several studies that indicated improved outcomes in accuracy when automated techniques were used for subject indexing in scientific fields.[2] In the same article, however, Golub remarked in reference to fully automated subject indexing approaches that "algorithms are really not able to entirely replace the intellectual work of subject indexing professionals."[3] It is also apparent that studying the way subject indexing professionals would utilize automated services in practice needs sustained scholarly inquiry to parallel technical advances.[4] Scholarship centered on knowledge worker preference with respect to automated machine learning support and artificial intelligence (AI) has found that "AI in knowledge work needs to focus not on full automation but rather on collaborative approaches where humans and AI work closely together," and that "boundaries, factors and circumstances of such collaboration should be studied empirically."[5]

The empirical project herein utilizes Annif machine learning software as an example case for professionals to consider features of machine learning models. The Annif automated subject indexing approach requires loading a linked data subject vocabulary first, and then processing training data for that vocabulary. System designers may select one or several (e.g., ensemble) algorithms for the model. The Annif system then generates subject suggestions in the targeted linked data vocabulary. Previous work explored methods for semi-automated subject suggestions using Annif, introducing data flows and likely feedback mechanisms to develop semi-automated support.[6] Prior work on Annif development for use within linked data editors concluded that "If automation is to be useful for the communities it

Jim Hahn (jimhahn@upenn.edu) is Head of Metadata Research, University of Pennsylvania Libraries, and PhD Student, School of Information Sciences, University of Illinois at Urbana-Champaign.

Library Resources & Technical Services | July 2024
https://doi.org/10.5860/lrts.68n3.8066

seeks to support, it must be ushered in with profound appreciation for, and in collaboration with, the professionals the automation would support."[7]

Compiling user feedback on semi-automated support as a departure point shifts from generalized subject support into domain-specific areas. The reference herein to domain is to that of a linguistic domain (e.g., semantics and syntax of a metadata description), not a domain of study (e.g., engineering, or medical fields). Blair reviewed seminal work pertaining to information retrieval and the philosophy of language; particularly the philosophers of language that argued for contextual uses of language qua meaning.[8] Blair goes on to extend this into document descriptions: "If the contexts of activities and practices are important for understanding language, it stands to reason that activities and practices are important for understanding document descriptions, too."[9] Hence, the metadata description as linguistic domain and object of analysis.

This work continued a sustained inquiry into the Annif software tool for semi-automated subject-indexing support.[10] Linked data editors can integrate Annif by way of common web development patterns. This approach represents advances over what was possible in automating subject indexing previously—prior work underscored how challenging the Library of Congress Subject Headings (LCSH) vocabulary was and indicated that it was too complex to enable automated services useful to libraries.[11] Indeed, due to certain systematic idiosyncrasies in LCSH, prior research found that, "the automatic analysis on syntactic structures in LCSH failed to uncover the precise semantics intended in subject headings in all cases because of their innate inconsistency. Introducing predictable syntaxes into LCSH and using them consistently will greatly assist in mining correct semantics of subject headings, predictable in that intended semantics can be retrievable based on the syntax."[12] More recent scholarship underscores the importance of formalizing subject languages for linked data applications, e.g., the semantic web: "The LCSH construction mechanism manages semantics very thinly in terms of its formalistic representations, while using natural language extensively for enhancing its flexibility and expressiveness; however, this expressiveness cannot be transferred to the semantic web."[13]

The atomic parts of subject languages are an exemplar of the attributes found in metadata descriptions. The database literature defines attributes as, "some property of interest that further describes an entity."[14] The *Intellectual Foundation of Information Organization* includes a chapter on bibliographic languages, and within the scope of this chapter posits the following exemplars of attributes: "author, title, edition, and subject."[15] The properties of metadata descriptions are the general attributes that comprise bibliographic descriptions. The mathematician Vladimir Arnold, whose work was applied to areas of non-linear systems, wrote that bifurcation is a "branching process" that "is widely used to describe any situation in which the qualitative, topological picture of the object we are studying alters with a change of the parameters on which the object depends."[16] To borrow from Arnold, the uses of metadata attributes here represent the parameters which, it is theorized, are influential in the way subject terminology is assigned. If the training data branch according to attributes in the metadata description, the results are a new set of training data, with qualitatively different parameters than non-bifurcated or general sets of training data. The two overarching research question addressed in this study are: (1) how can knowledge workers be involved in curating and selecting the data used as

the input in machine learning and (2) does knowledge worker collaboration in machine learning model development lead to acceptance and use of machine learning tools by knowledge workers? A survey developed for this research inquired about the types of ongoing feedback in which professionals may be interested in participating to shape future semi-automated service development.

The background section details the linguistic domain areas of language and genre, which provides useful context for the methods and results sections that follow. The paper concludes with a discussion of the key survey findings and a comparative analysis of pilot Annif services developed with the input from the survey. The pilot Annif services include general Annif machine learning models contrasted with models that are branched according to genre preferences found in the survey. The discussion will delineate subject retrieval metrics for specialized Annif genre services contrasted to non-bifurcated (or general) Annif services. Results showed that several genres of interest were improved by utilizing genre focused training data.

Background

With respect to language branching in the Annif set of services, there are ample approaches to the treatment of language in subject indexing in the information science literature. The "Multilingual Access to Subject" or MACS project contains human-derived mapping of vocabularies.[17] In the MACS project, library professionals were engaged in mapping among different subject heading languages to allow searchers to use their preferred languages to search. The strategy employed in this linking was human-derived mappings among German, French, and English subject vocabularies.[18] This approach contrasts with "direct bilingual or multilingual subject cataloging policies and practices."[19] Another example of support for languages is the Bibliotheca Alexandrina's Subject Authority File and Linked Subject Data where users who search Arabic titles can search in non-Romanized Arabic scripts.[20]

In the context of the Annif machine learning software, support for languages has evolved over time. Previous versions of the Annif software referenced a language code in the source vocabulary. However, since the release of the 0.59 version, the software no longer requires a specific language in the source vocabulary; in effect, the Annif system is now multilingual.[21] An out-of-distribution problem is the result of training data that are unlike the data used to make a prediction in the real-word situation.[22] Attending to the linguistic attributes of metadata descriptions, including genre/form type, may help to ensure that the attributes in a metadata description improve the precision and recall of a subject indexing term and ultimately prevents out-of-distribution errors.

The development of genre specific Annif services trained only on subjects from a targeted genre was tested in this paper. These genre-specific targets may provide more accurate subject indexing terms. To develop such a service the training data included only subjects assigned in a metadata description that was from a targeted genre domain. A previous conference paper detailed the challenge of re-using the genre/form assignments.[23] In particular, the work by Lee and Zhang reported "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."[24] The authors noted further in their study

that genre (terms) "are forms of social action."[25] Hider, White, and Barlow explored uses of genre in domains outside of libraries and made a systematic study of mapping genre terms from other domains (such as the International Movie Database, Wikipedia, and others) to the Library of Congress Genre Form Terms (LCGFT).[26] The Library of Congress notes that LCGFT is a distinct vocabulary and that,

> Genres and forms may be broadly defined as categories of resources that share known conventions. More specifically, genre/form terms may describe the purpose, structure, content, and/or themes of resources. Genre/form terms describing content and themes most frequently refer to creative works and denote common rhetorical devices that usually combine elements such as plot and setting, character types, etc. Such terms may be closely related to the subjects of the creative works but are distinct from them.[27]

The findings from Hider, White, and Barlow showed that "some vocabularies of film genre are considerably closer to that of the LCGFT than are others, but that overall alignment between film vocabularies is hard to predict and dependent on a number of interrelated variables."[28] The authors noted that differing perspectives of genre are an important consideration when mapping among varied vocabularies.

More recently, scholarship on establishing a framework for formats has been successful in providing a structure to understand and address elements of format that have become entwined with genre types.[29] Specifically, a binary was introduced among the conceptualization of containers to genres; examples of a container were asserted to be books and journals, whereas exemplars of genre included articles, chapters, and editorials.[30] The conceptualization of genres and container types as abstractions bears a striking resemblance to the manner in which the concept of works in library and information science are frequently referenced: "genres are abstractions," and "they should be understood as mental models that people develop and deploy when trying to achieve certain types of genre actions," and, "container types are also abstractions and an individual publication is an instantiation of a container type."[31] This abstraction for genres and containers is a valuable frame which this present research will turn to again in the discussion of key findings.

It is important to underscore, however, that in the case of both the MACS project on languages and the LCGFT mapping study of genres, human intermediaries were involved in curating these mappings. Mappings among disparate vocabularies are a foundation of metadata interoperability.[32] Professionals with subject expertise can improve the labeling of small amounts of data to improve certain types of machine learning projects when large sets of data are not available.[33] A key consideration in this study is in understanding professionals' preferred vocabularies, genres, and languages to include in semi-automated subject indexing.

Methods

An internet-based survey targeted catalogers from linked data domains. The Annif machine learning software uses linked data vocabularies for base vocabularies and training data. Because the target vocabularies are in linked data, participants from the linked data domains were sought for the study.

A questionnaire gathered cataloger perceptions and beliefs for targets of machine learning that could support semi-automated indexing.[34] The LD4 Community was a source of survey advertisement and outreach. This served the interest of the research in the sense that the outputs that Annif generate are from linked data vocabularies. According to the LD4 Community Charter, "LD4 is a community that works together to advance library and archival practices. We focus on linking and using data on the Web to advance the mission, goals and objectives of libraries and archives."[35] Additional promotion of the survey was made on the metadata librarians' listserv and through the professional social networking site LinkedIn.[36]

With this purposeful sampling method, there was not an attempt to obtain generalizable findings. Rather, the survey is meant to inform development of machine learning services. As the survey sought to understand preferences of semi-automated indexing system features, no attempt to discern or otherwise gather the demographic information of participants was undertaken.

The survey was available throughout October 2022 and nearly seventy individuals participated. Respondents had the option to skip questions of the questionnaire. Survey questions were informed by the broad research concerns: what are the types of cataloging that could make use of semi-automated technology for subject assignment, how the users of such a system might expect it to behave, and how those users believe they should be consulted in the development of semi-automated support. In asking about the controlled vocabularies, the world languages, and the variety of possible genres that semi-automated subject indexing may support, the survey questions are directly addressing the perceptions and beliefs of catalogers working in linked data description.

The author did not obtain institutional review board (IRB) approval for the study. The IRB office has classed this type of research within the University of Pennsylvania Libraries as quality improvement when no demographic data are collected about human participants. More specifically, when no information on the participants themselves are collected, the IRB office at University of Pennsylvania would consider this to qualify as non-human subject research. The overall focus of the study is evaluation of the subject indexing machine learning system results and not evaluation of catalogers working in linked data. Where survey results from catalogers are reported, these are analyzed in the aggregate.

Results

The perceived usefulness of a semi-automated subject suggestion tool was the first topic of the survey. Table 1 delineates responses to the statement, "Semi-automated support may be useful in my work if I can shape the service through feedback."

With respect to the question about vocabulary targets for semi-automated support, there was a large percentage for LCSH and some interest as well in Faceted Application of Subject Terminology (FAST) and Medical Subject Headings (MeSH), although to a lesser extent.

Table 1. Responses to the statement: Semi-automated support may be useful in my work if I can shape the service through feedback.

Choice	Count	Percent of Data	Confidence Interval (Percent of Data)
Strongly agree	33	48.5	37.1 to 60.2
Somewhat agree	25	36.8	26.3 to 48.6
Neither agree nor disagree	5	7.4	3.2 to 16.1
Somewhat disagree	3	4.4	1.5 to 12.2
Strongly disagree	2	2.9	0.8 to 10.1

Table 2. Responses for the question: Which vocabularies would be most amenable to semi-automated suggestions for incorporation in your work?

Choice	Checked Percent	Confidence Interval	Checked Count	Sample Size
LCSH	82.0	70.5 to 89.6	50	61
FAST	52.5	40.2 to 64.5	32	61
Other	34.4	23.7 to 47.0	21	61
MeSH	21.3	12.9 to 33.1	13	61

Responses in the "Other" category included mentions of the following vocabularies: Homosaurus;[37] Getty Art & Architecture Thesaurus;[38] Conspectus; Répertoire de vedettes-matière de l'Université Laval (RVM);[39] The Virtual International Authority File (VIAF);[40] Library of Congress Demographic Group Terms;[41] Library of Congress Medium of Performance Thesaurus;[42] Gender, Sex, and Sexual Orientation (GSSO) ontology;[43] Rare Books and Manuscripts Section (RBMS) Controlled Vocabularies;[44] Library of Congress Children and Young Adults Cataloging;[45] Canadian name authority & subject headings;[46] National Library of Israel;[47] Thesaurus for Graphic Materials,[48] and Geonames.[49]

The survey included a prompt for the targeted language desired to support subject suggestions. Table 3 shows the top twenty responses for languages support.

Nearly 30 percent of the responses in the language category included "Other" languages with suggestions for the following: Spanish; Yiddish; Ukrainian; Belarusian; Serbian; Croatian; any of the Slavic and East European Languages, and Modern Turkish. Table 4 delineates the responses to which genre areas would be most useful for semi-automated subject support.

A total of 17 percent of the responses to the genre form question selected the "Other" category for genre form and reported the following areas as potential targets: catalogs at exhibitions; belle lettres; censuses; local history; electronic books; graphic materials (posters, postcards, photographs, etc.), and religion. The survey provided an option for genre as a combination of those terms found in FAST and LCGFT (e.g., Biography in FAST, Biographies—LCGFT). Table 5 shows the responses to the question, "If Annif could provide results based upon a selected country of origin for a publication, please register which countries you would like to be able to select as a focus of semi-automated suggestions."

Table 3. Top 20 responses: If you could select a targeted language support for semi-automated subject suggestions, select those languages that would be most useful to your work.

Choice	Checked Percent	Confidence Interval	Checked Count	Sample Size
English	76.5	63.2 to 86.0	39	51
French	35.3	23.6 to 49.0	18	51
German	31.4	20.3 to 45.0	16	51
Other	29.4	18.7 to 43.0	15	51
Russian	19.6	11.0 to 32.5	10	51
Hebrew	17.6	9.6 to 30.3	9	51
Italian	15.7	8.2 to 28.0	8	51
Korean	15.7	8.2 to 28.0	8	51
Chinese	13.7	6.8 to 25.7	7	51
Arabic	11.8	5.5 to 23.4	6	51
Japanese	11.8	5.5 to 23.4	6	51
Portuguese	9.8	4.3 to 21.0	5	51
Czech	7.8	3.1 to 18.5	4	51
Greek, Modern (1453-)	7.8	3.1 to 18.5	4	51
Hungarian	7.8	3.1 to 18.5	4	51
Polish	7.8	3.1 to 18.5	4	51
Bosnian	5.9	2.0 to 15.9	3	51
Bulgarian	5.9	2.0 to 15.9	3	51
Dutch	5.9	2.0 to 15.9	3	51
Romanian	5.9	2.0 to 15.9	3	51

The survey also asked whether the respondent would recommend the Annif service to a friend or colleague, as word of mouth recommendations may be representative of trust.[50] Table 6 contains the responses.

The survey asked catalogers whether semi-automated support might be useful in their work if they can modify the service through feedback. Not all of those who took the survey responded to this question about ongoing feedback. Table 7 delineates the types of continued feedback that catalogers would be interested in participating.

Discussion

This discussion section comprises three parts. First, the limitations of the study are discussed so that survey analysis can be contextualized with an understanding of those limits. A survey analysis follows, with an analysis and interpretation of the survey results. Following the survey analysis, the discussion then transitions into an empirical test of how machine learning models trained on genre data contrasts

Table 4. Genre form responses.

Choice	Checked Percent	Confidence Interval	Checked Count	Sample Size
History	63.8	49.5 to 76.0	30	47
Biography	44.7	31.4 to 58.8	21	47
Fiction	40.4	27.6 to 54.7	19	47
Handbooks, manuals, etc.	34.0	22.2 to 48.3	16	47
Biographies	31.9	20.4 to 46.2	15	47
Academic theses	31.9	20.4 to 46.2	15	47
Criticism, interpretation, etc.	29.8	18.7 to 44.0	14	47
Poetry	29.8	18.7 to 44.0	14	47
Pictorial works	29.8	18.7 to 44.0	14	47
Periodicals	27.7	16.9 to 41.8	13	47
Conference papers and proceedings	27.7	16.9 to 41.8	13	47
Sound recordings	27.7	16.9 to 41.8	13	47
Scores	27.7	16.9 to 41.8	13	47
Congresses	25.5	15.3 to 39.5	12	47
Bibliographies	23.4	13.6 to 37.2	11	47
Catalogs	23.4	13.6 to 37.2	11	47
Dictionaries	23.4	13.6 to 37.2	11	47
Video recordings	23.4	13.6 to 37.2	11	47
Electronic journals	21.3	12.0 to 34.9	10	47
Early works to 1800	19.1	10.4 to 32.5	9	47
Maps	19.1	10.4 to 32.5	9	47
Drama	19.1	10.4 to 32.5	9	47
Bibliography	17.0	8.9 to 30.1	8	47
Other	17.0	8.9 to 30.1	8	47
Exhibitions	12.8	6.0 to 25.2	6	47
Internet videos	12.8	6.0 to 25.2	6	47
Statistics	10.6	4.6 to 22.6	5	47
Sources	8.5	3.4 to 19.9	4	47
Early works	8.5	3.4 to 19.9	4	47
Private bills	2.1	0.4 to 11.1	1	47

with the machine learning models that do not branch based on genre specific attributes. The empirical evaluation in the third part of this discussion reports a comparison of precision and recall metrics for tests of bifurcated and non-bifurcated machine learning approaches.

A limitation for generalization of survey results include the use of a purposeful sampling method which sought to target catalogers working with linked data and is not representative of all professionals who

Table 5. Top 20 results on country of publication.

Choice	Checked Percent	Confidence Interval	Checked Count	Sample Size
United States of America	73.7	58.0 to 85.0	28	38
United Kingdom of Great Britain and Northern Ireland	28.9	17.0 to 44.8	11	38
France	23.7	13.0 to 39.2	9	38
Germany	23.7	13.0 to 39.2	9	38
Spain	23.7	13.0 to 39.2	9	38
Canada	18.4	9.2 to 33.4	7	38
Mexico	18.4	9.2 to 33.4	7	38
Israel	15.8	7.4 to 30.4	6	38
Russian Federation	15.8	7.4 to 30.4	6	38
Brazil	13.2	5.8 to 27.3	5	38
Colombia	13.2	5.8 to 27.3	5	38
Ecuador	13.2	5.8 to 27.3	5	38
Peru	13.2	5.8 to 27.3	5	38
Ukraine	13.2	5.8 to 27.3	5	38
China	10.5	4.2 to 24.1	4	38
Cuba	10.5	4.2 to 24.1	4	38
Czech Republic	10.5	4.2 to 24.1	4	38
Italy	10.5	4.2 to 24.1	4	38
Romania	10.5	4.2 to 24.1	4	38
Serbia	10.5	4.2 to 24.1	4	38

Table 6. Recommend Annif to a friend or a colleague?

Choice	Count	Percent of Data	Confidence Interval (Percent of Data)
Yes	7	17.1	8.5 to 31.3
No	7	17.1	8.5 to 31.3
Maybe	27	65.9	50.5 to 78.4

do knowledge work in libraries. Nonetheless, purposeful sampling methods address the aims of the study, which are to understand the broad question of professionals involved in curating and selecting the data used as input into machine learning and its perceived usefulness. Outside of generalization limitations, a more specific limitation is that soliciting survey data from the Internet is a trade off between reaching more people, and obtaining deep qualitative interview data from an in-person interview that offers a chance for follow up and capture of body language and sentiment. Future systems that are built could be tested on small sets of users in an in-person study to supplement the survey data from this study.

Table 7. Feedback preferences.

Choice	Checked Percent	Confidence Interval	Checked Count	Sample Size
Virtual workshop and discussion	59.0	43.4 to 72.9	23	39
Not interested	38.5	24.9 to 54.1	15	39
In-person focus group feedback	7.7	2.7 to 20.3	3	39
Other	2.6	0.5 to 13.2	1	39
In-person workshop and discussion	0.0	0.0 to 9.0	0	39

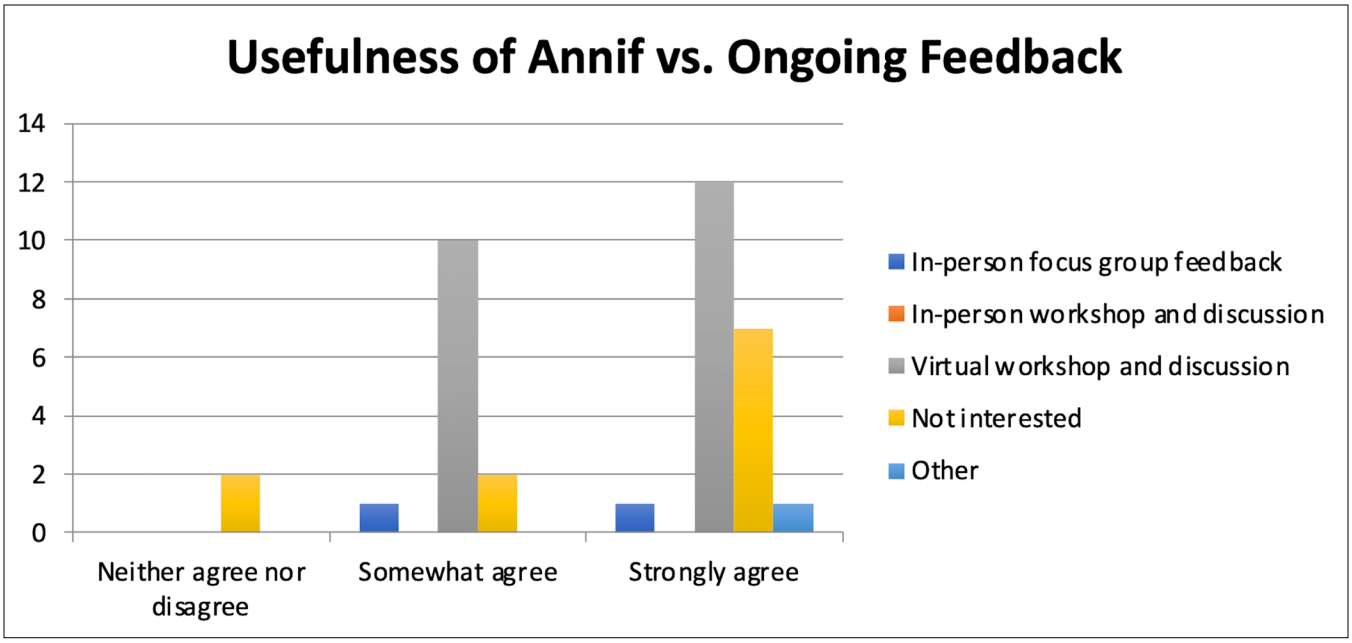

Figure 1. Comparison of responses of those with neutral or positive expectations that Annif would be helpful in their work if modified through ongoing feedback contrasted with the types of ongoing feedback.

The responses to the research question on the usefulness of semi-automated support if shaped through feedback, shown in table 1, indicated that 48.5 percent of those who responded to this question strongly agree, 38.6 percent somewhat agree, 7.4 percent neither agree nor disagree, and 4.4 percent somewhat disagree, and 2.9 percent strongly disagree. With respect to modifying data inputs for machine learning service with the notions of ongoing feedback, table 7 shows that nearly 60 percent of the respondents favored providing ongoing feedback through virtual workshops or discussions. Perhaps not very surprisingly, those who wanted to provide feedback in virtual workshops or online discussions appeared to be those in groups that may find Annif service useful if they are able to modify it. It is unclear why a large majority of those uninterested in providing additional feedback (beyond this survey) were from the group of respondents who registered initial "strong agreement" to the question of the usefulness of modifying a service through feedback (survey question 1). One possibility may be attributable to the generalized nature of the request, e.g., "when the language is too vague or generic, it gives respondents little to no information on what they would expect and hence lower their willingness."[51] Another possible explanation for this incongruity in the data may be a result of survey fatigue, where attention to

the questions diminishes over the course of a survey.[52] The analysis of the respondent groups is shown in figure 1.

Concerns remain for large-scale automation of subject indexing. A report on using automated methods in software engineering found serious issues with machine learning models that may perpetuate existing biases in training data.[53] Approaches to highlight bias must include sufficient understanding of the datasets used in the training and the outputs of the trained model. Further, there are ample examples of ethical principles to address in AI guidelines.[54] Recently, the US Department of Commerce National Institute of Standards and Technology (NIST) released its "Artificial Intelligence Risk Management Framework,"[55] which is described as a "guidance document for voluntary use by organizations designing, developing, deploying or using AI systems to help manage the many risks of AI technologies."[56] The release of a framework from a national standards body suggests both the promise and the significant risks of implementing AI technologies.

The responses to the more specific research questions in the survey about the types of data input into machine learning showed a wide range of catalogers' preferences in vocabulary, language, and genre. Beyond the survey prompt of LCSH, FAST, and MeSH, the "Other" category, which was a free text field, allowed survey respondents to input their suggestions. As table 2 shows, the results included 34.4 percent additional vocabularies, or higher than the MeSH, which 21.3 percent of the sixty-one survey responses in this area selected. LCSH was the top vocabulary that 82 percent of the respondents selected, followed by FAST as the second most popular that 52.5 percent of the respondents chose.

There was a variety of possibilities for languages–the top languages reported included English, French, and German. A drawback of the survey is that the responses represent only those who were able to access the survey in English. As with the previous question on vocabulary targets, 29.4 percent—or fifteen of the fifty-one responses—checked the "Other" category for language. History (63.8 percent), Biography (44.7 percent), and Fiction (40.4 percent) were the top responses within the genre form inquiry.

With respect to service set development for semi-automation, it appears LCSH and FAST would be the most popular areas used and that there may be many users of an Annif service within the English, French, and German languages. The LCGFT implementation in Annif could also progress with a similar strategy for languages, as it may be employed if there are ample training data of sufficient quality.

In the context of genre type preferences that were reported in the survey, a subsequent test of the viability of branching along these lines was undertaken. The evaluation is both computational, in measures of retrieval metrics, as well as comparative. The comparative tasks are to evaluate the performance of Annif models which are trained only on genre data to general Annif projects which use no branching of training data for genre. The survey results for genre are used as model bifurcations in the subsection which follows.

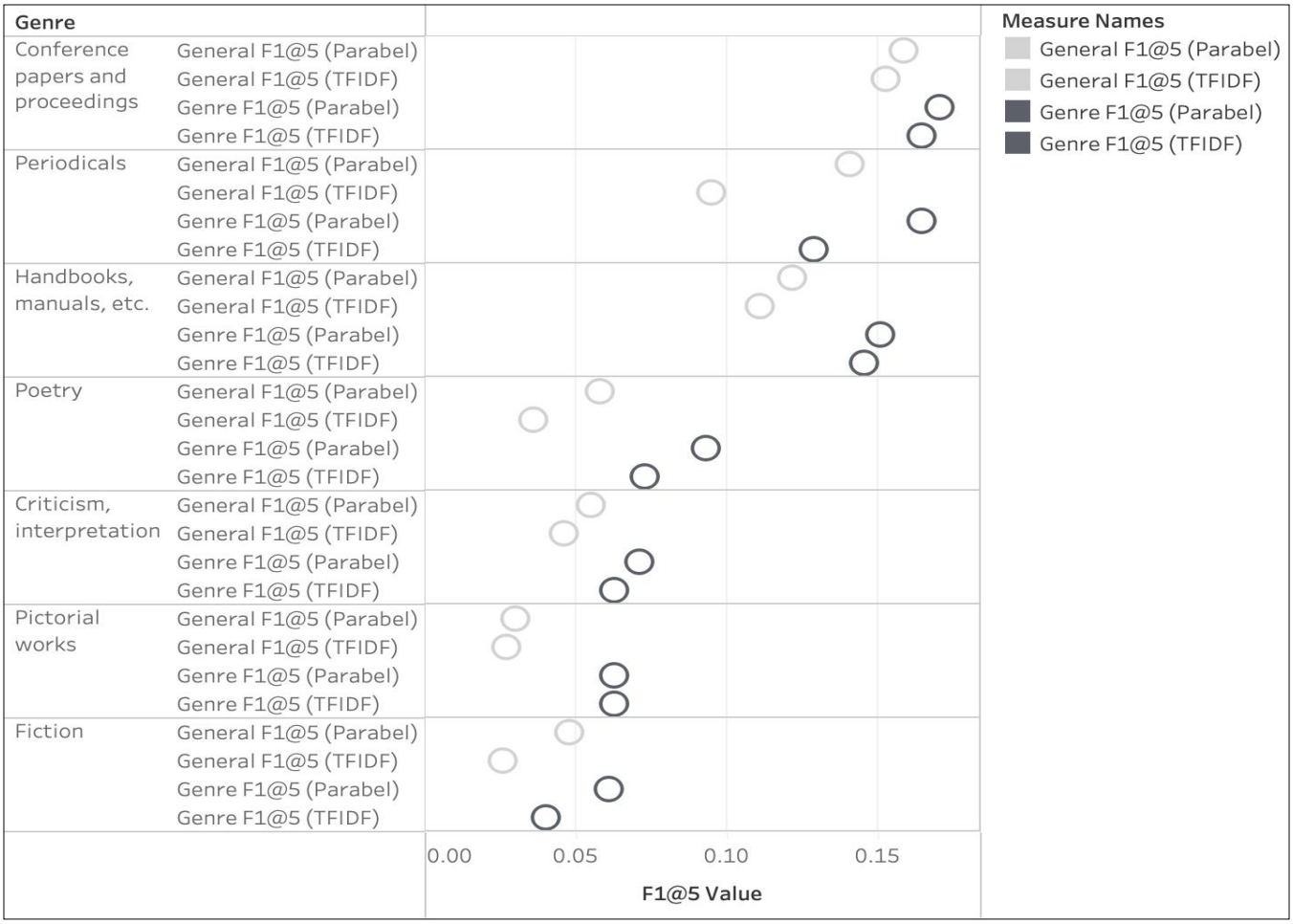

Figure 2. Comparison of Annif Algorithms applied to a General (Non-bifurcated) Corpus and a Bifurcated Genre Corpus–using LCSH Linked Data Vocabulary as the base vocabulary.

Survey Results as Pilot Bifurcations in Annif Services

The perspective of this paper is that bifurcating machine learning services may result in better precision and recall for subject assignment. Genre results from the survey were used to inform bifurcating machine learning models in Annif. These pilot tests seek to address the questions: how do bifurcated machine learning models perform as compared to general models with no bifurcation of genre—and does the algorithm used have any influence on the scoring outcome? Figure 2 shows a comparison of the F1 scores among general and bifurcated Annif services. The F1 score is an evaluation metric in machine learning projects—it is a measure that combines recall and precision metrics, more formally it is referred to as the harmonic mean of precision and recall.[57] To reproduce the Annif evaluations please consult the Zenodo open repository.[58]

The analysis compared two non-bifurcated Annif services for ten Annif bifurcations using test sets for the genres. The algorithms used in testing included the tf-idf (term frequency and inverse document frequency, an algorithm that makes a prediction over the importance of a term by its presence in an

Table 9. LCSH Category Semantics: five major classes of terms.

Term Class	Purpose
Main or focal headings	Functioning as lead terms in subject headings, these terms are used to denote the essential aboutness of documents being described.
Topical subheadings	The purpose of these terms is to qualify main headings and subheadings.
Form or document types	Used for the purpose of qualification.
Chronological periods	Used for qualification.
Geographical areas	Used for qualification.

Source: Elaine Svenonius, "LCSH: Semantics, Syntax and Specificity," *Cataloging & Classification Quarterly* 29, no. 1–2 (2000): 17–30, https://doi.org/10.1300/J104v29n01_02.

index of the document terms)[59] and the Parabel algorithm.[60] The Scikit-Learn python method for splitting training and testing data held back a randomized 20 percent of the data for use in testing.[61]

Shown in figure 2 are the results of F1 scores; genre Annif services improved for seven of the ten services. Genres with improved scores over general subject indexing included Fiction, Handbooks, Manuals, etc., Criticism, interpretation, Poetry, Pictorial works, Periodicals, and Conference proceedings. When a linked data editor has context information about the genre form in certain cases the genre specific Annif service is more accurate than general models that do not address genre/form.

What do the genre forms with higher scores have in common—is there a common trait among those models with the highest genre scores? A linguistic framing is a useful formalism, because retrieval "of documents or textual material—is fundamentally a linguistic process."[62] The linguistic structure of LCSH include the syntax and semantics—leaving aside for the moment the important, though out of scope (for this paper), content of the LCSH labels.[63] The linguistic analysis undertaken by Svenonius analyzed the category semantics shown in table 9.

Svenonius reported the three most common syntactic constructions in LCSH contain form qualification.[64] For those genres in figure 2 that have higher scoring in the genre only models as compared to general, non-bifurcated models, it seems that machine learning model construction can build an association among the qualifiers for form or document types (delineated as term class 3 in table 9).

Conclusion

This paper reported survey results of knowledge workers' preferences for machine learning-based support, and the findings showed their desire for continued consultation and feedback through virtual meetings and workshops. The paper also investigated both acceptance and use of machine learning tools. Emerging results indicated that such services would be useful if catalogers were able to shape the service; nearly half of those responding to this question strongly agreed to this proposition.

Future virtual workshops—the preferred method to gather continued input—will provide an opportunity to inform data inputs and further uses of machine learning models used in semi-automated

subject indexing. Future research will report how machine learning models evolve and will analyze the effects of the machine learning model changes in an Annif system. Follow up studies will report on community-supported machine learning models in further detail.

Bifurcation approaches were informed by the survey of knowledge worker preferences for vocabulary, language, and genre. Tests of these bifurcations improved retrieval metrics of automated subject indexing for several, but not all genres. By testing the hypothesis that training on bifurcated sets of genre data as compared to general (non-bifurcated) machine learning services, this research provides empirical and reproducible evidence for the advantages gained by bifurcation paths with respect to genre. Future research will test additional language-based bifurcations for semi-automated subject indexing.

Longer-term sustained inquiry into the problem of semi-automated subject indexing may include considerations for the way metadata descriptions depend upon conceptualizations of genre with respect to social and historical context. Catalogers in the late twentieth century may have thought about genre in different ways from the way catalogers describe genres in the twenty-first. Social rules governing metadata descriptions are important to understanding the context of description.[66] Levinson asserted that "musical composition could not fail to be seen as a historically rooted activity whose products must be understood with reference to their points of origin."[66] Can this perspective on musical composition be applied to considerations of genres used in bibliographic description? In *Work & Object,* Peter Lamarque stated that works "are cultural artefacts among whose essential properties are intentional or relational properties (i.e. that what they are as works is partially dependent on how they are taken to be by qualified observers)."[67] These intellectual foundations may inform a perspective wherein the creation of genre may be viewed as a social activity within a contextual cultural moment. The implications for semi-automated indexing are that as much as new technologies advance, they too must reference and adhere to the social context (and social rules) which technology seeks to support. In another sense—new indexing techniques must acknowledge our old indexing cultures, while simultaneously co-creating new cultures of description that may be an ever more collaborative endeavor of knowledge machines and knowledge professionals.

Referencess

1. David A. Mindell, *Our Robots, Ourselves: Robotics and the Myths of Autonomy* (New York: Viking, 2015).
2. Koraljka Golub, "Automated Subject Indexing: An Overview," *Cataloging & Classification Quarterly* 59, no. 8 (2021): 702–19, https://doi.org/10.1080/01639374.2021.2012311.
3. Golub, "Automated Subject Indexing," 703.
4. Jung-ran Park and Caimei Lu, "Application of Semi-automatic Metadata Generation in Libraries: Types, Tools, and Techniques," *Library & Information Science Research* 31, no. 4 (2009): 225–31, https://doi.org/10.1016/j.lisr.2009.05.002.
5. Konrad Sowa, Aleksandra Przegalinska, and Leon Ciechanowski, "Cobots in Knowledge Work," *Journal of Business Research* 135–42, (2021): 135, 141, https://doi.org/10.1016/j.jbusres.2020.11.038.

6. Jim Hahn, "Semi-automated Methods for BIBFRAME Work Entity Description," *Cataloging & Classification Quarterly* 59, no. 8 (2021): 853–67, https://doi.org/10.1080/01639374.2021.2014011.
7. Hahn, "Semi-automated Methods," 864.
8. Ludwig Wittgenstein, *Philosophical Investigations* (New York: Macmillan, 1953); Hilary Putnam, *Representation and Reality,* (Cambridge, MA: MIT Press, 1991); Peter Frederick Strawson, "On Referring," *Mind,* 59 (1950): 320–44, https://www.jstor.org/stable/2251176.
9. David C. Blair, "Information Retrieval and the Philosophy of Language," *Annual Review of Information Science and Technology* 37 (2003): 3–50, https://doi.org/10.1002/aris.1440370102.
10. Osman Suominen, "Annif: DIY Automated Subject Indexing Using Multiple Algorithms," *LIBER Quarterly* 29, no. 1 (2019): 1–25, https://doi.org/10.18352/lq.10285; Osman Suominen, Juho Inkinen, and Mona Lehtinen, "Annif and Finto AI: Developing and Implementing Automated Subject Indexing," *JLIS.It* 13, no. 1 (2022): 265–82, https://www.jlis.it/index.php/jlis/article/view/437.
11. Jung-ran Park and Andrew Brenza, "Evaluation of Semi-automatic Metadata Generation Tools: A Survey of the Current State of the Art," *Information Technology and Libraries* 34, no. 3 (September 2015): 22–42, https://doi.org/10.6017/ital.v34i3.5889.
12. Kwan Yi and Lois Mai Chan, "Revisiting the Syntactical and Structural Analysis of Library of Congress Subject Headings for the Digital Environment," *Journal of the American Society for Information Science and Technology* 61 (2010): 677–87, https://doi.org/10.1002/asi.21295.
13. Manolis Peponakis et al., "Expressiveness and Machine Processability of Knowledge Organization Systems (KOS): An Analysis of Concepts and Relations," *International Journal on Digital Libraries* 20 (2019): 433–52, https://doi.org/10.1007/s00799-019-00269-0.
14. Elmasri Ramez and Sham Navathe, *Fundamentals of Database Systems* (Hoboken, NJ: Pearson, 2020), 33.
15. Elaine Svenonius, *The Intellectual Foundation of Information Organization* (Cambridge, MA: MIT Press, 2000), 53.
16. Vladimir I. Arnold, "Lectures on Bifurcations in Versal Families," in, Alexander B. Givental et al., eds., *Vladimir I. Arnold—Collected Works (Vol 2)* (Berlin, Heidelberg: Springer, 1972), https://doi.org/10 .1007/978-3-642-31031-7_29, 54.
17. Patrice Landry, "Multilingual Subject Access: The Linking Approach of MACS," *Cataloging & Classification Quarterly* 37, no. 3–4 (2004): 177–91, https://doi.org/10.1300/J104v37n03_11.
18. German National Library, Linked Vocabularies (Voclink), "MACS: Multilingual Access to Subjects," updated July 30, 2020, https://www.dnb.de/EN/voclink.
19. Landry, "Multilingual Subject Access," 179.
20. Magda El-Sherbini, "Multilingual subject retrieval—Bibliotheca Alexandrina's Subject Authority File and Linked Subject Data," in *Data Science, Learning by Latent Structures, and Knowledge Discovery,* Berthold Lausen, Sabine Krolak-Schwerdt, and Matthias Böhmer, eds. (Springer, 2015), http://www.springer.com /statistics/book/978-3-662-44982-0.
21. Osma Suominen et al., "Annif," Zenodo, September 23, 2022, http://doi.org/10.5281/zenodo.7107271.

22. Dan Hendrycks and Kevin Gimpel "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks," *5th International Conference on Learning Representations, ICLR 2017—Conference Track Proceedings* (January 2017), https://arxiv.org/abs/1610.02136.

23. Jim Hahn, "Cataloger Acceptance and Use of Semiautomated Subject Recommendations for Web Scale Linked Data Systems," 87th IFLA World Library and Information Congress (WLIC)/2022 in Dublin, Ireland; International Federation of Library Associations and Institutions (IFLA), https://repository.ifla.org/handle/123456789/1955.

24. Hur-Li Lee and Lei Zhang, "Tracing the Conceptions and Treatment of Genre in Anglo-American Cataloging," *Cataloging & Classification Quarterly* 51, no. 8 (2013): 909, https://doi.org/10.1080/01639374.2013.832457.

25. Lee and Zhang, "Tracing the Conceptions," 892.

26. Philip Hider, Hollie White, and Phillipa Barlow, "Film Genres through Different Lenses: Mapping Commonly used Film Vocabularies onto the Library of Congress Genre/Form terms," *Library Trends* 69, no. 3 (2021): 630–45, http://doi.org/10.1353/lib.2021.0007.

27. Library of Congress, Introduction to Library of Congress Genre/Form Terms for Library and Archival Materials, 2022 edition, https://www.loc.gov/aba/publications/FreeLCGFT/2022-LCGFT-intro.pdf, p. 3.

28. Library of Congress, Introduction, 643.

29. Brittany Brannon et al., "Genre Containers: Building a Theoretical Framework for Studying Formats in Information Behavior," *Journal of the Association for Information Science and Technology* 73, no. 4 (2022): 609–24, https://doi.org/10.1002/asi.24600.

30. Brannon et al., "Genre Containers," 614.

31. Brannon et al., "Genre Containers," 619.

32. Bernhard Haslhofer and Wolfgang Klas, "A Survey of Techniques for Achieving Metadata Interoperability," *ACM Computing Survey* 42, no. 2 (February 2010), https://doi.org/10.1145/1667062.1667064

33. Kanyao Han et al., "An Expert-in-the-Loop Method for Domain-Specific Document Categorization Based on Small Training Data," *Journal of the Association for Information Science and Technology* (2022): 1–16, https://doi.org/10.1002/asi.24714.

34. Lynn Silipigni Connaway and Marie L. Radford. *Research Methods in Library and Information Science, 6th Edition*, Vol. Library and Information Science Text Series (Santa Barbara, CA: Libraries Unlimited, 2017), 153.

35. LD4 Community, accessed September 3, 2023, https://sites.google.com/stanford.edu/ld4-community-site/home .

36. metadataLibrarians listserv, accessed September 17, 2023, http://metadatalibrarians.monarchos.com/; Cataloger Feedback Requested: Semi-automation, accessed September 17, 2023, https://www.linkedin.com/pulse/cataloger-feedback-requested-semi-automation-jim-hahn.

37. Digital Transgender Archive, accessed November 23, 2022, https://homosaurus.org/releases.

38. Getty, accessed November 23, 2022, https://www.getty.edu/research/tools/vocabularies/aat/.

39. OCLC, Répertoire de vedettes-matière de l'Université Laval (RVM), accessed November 23, 2022, https://help.oclc.org/Metadata_Services/WorldShare_Record_Manager/Authority_records/Work_with_authority_records/RVM_Authorities?sl=en.

40. OCLC, The Virtual International Authority File, accessed November 23, 2022, https://viaf.org.

41. Library of Congress, Library of Congress Demographic Group Terms, accessed November 23, 2022, https://www.loc.gov/aba/publications/FreeLCDGT/freelcdgt.html.

42. Library of Congress, Library of Congress Medium of Performance Thesaurus for Music, accessed November 23, 2022, https://loc.gov/aba/publications/FreeLCMPT/freelcmpt.html.

43. Ontobee, Gender, Sex, and Sexual Orientation (GSSO) ontology, accessed November 23, 2022, https://ontobee.org/ontology/GSSO.

44. Rare Books and Manuscripts Section, RBMS Controlled Vocabularies, accessed November 23, 2022, https://rbms.info/vocabularies/introductions/GenreIntro.htm.

45. Library of Congress, LC Children and Young Adults Cataloging, accessed November 23, 2022, https://www.loc.gov/aba/cyac/faq.html.

46. OCLC, Canadian Name Authority & Subject Headings, accessed November 23, 2022, https://help.oclc.org/Metadata_Services/WorldShare_Record_Manager/Authority_records/Work_with_authority_records/Canadiana_Authorities.

47. National Library of Israel, accessed November 23, 2022, https://www.nli.org.il/en.

48. Thesaurus for Graphic Materials, accessed November 23, 2022, https://www.loc.gov/rr/print/tgm1/.

49. Geonames, accessed November 23, 2022, https://www.geonames.org/.

50. Bernard J. Jansen, Mimi Zhang, Kate Sobel, and Abdur Chowdury, "Twitter Power: Tweets as Electronic Word of Mouth," *Journal of the American Society for Information Science and Technology* 60 (2009): 2169–88, https://doi.org/10.1002/asi.21149.

51. Mingnan Liu, "Soliciting Email Addresses to Re-contact Online Survey Respondents: Results from Web Experiments," *Methodological Innovations* 13, no. 2 (2020): 7, https://doi.org/10.1177/2059799120937237.

52. Kylie Brosnan, Nazila Babakhani, and Sara Dolnicar, "'I Know What You're Going to Ask Me': Why Respondents Don't Read Survey Questions," *International Journal of Market Research* 61, no. 4 (2019): 366–79, https://doi.org/10.1177/1470785318821025.

53. Yuriy Brun, "The Promise and Perils of using Machine Learning When Engineering Software (keynote paper)," in *Proceedings of the 6th International Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE 2022)* (New York: Association for Computing Machinery, 2022), 1–4, https://doi.org/10.1145/3549034.3570200.

54. Anna Jobin, Marcello Ienca, and Effy Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1 (2019): 389–99, https://doi.org/10.1038/s42256-019-0088-2.

55. National Institute of Standards and Technologies, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," https://doi.org/10.6028/NIST.AI.100-1.

56. National Institute of Standard and Technologies, "NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence," updated January 26, 2023, https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial.

57. Leon Derczynski, "Complementarity, F-score, and NLP Evaluation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016),* 261–66.

58. Jim Hahn, "Data from: Bifurcation of Semi-Automated Subject Indexing Services (Version 1)," Zenodo, April 4, 2023, http://doi.org/10.5281/zenodo.7803233.
59. Akiko Aizawa, "An Information-Theoretic Perspective of TF–IDF Measures," *Information Processing & Management* 39, no 1 (2003): 45–65, https://doi.org/10.1016/S0306-4573(02)00021-3.
60. Yashoteja Prabhu et al., "Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising," in *Proceedings of the 2018 World Wide Web Conference* (2018): 993–1002.
61. Fabian Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825–30, https://dl.acm.org/doi/10.5555/1953048.2078195.
62. Blair, "Information Retrieval and the Philosophy of Language."
63. Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018).
64. Svenonius, "LCSH," 24.
65. John R. Searle, *The Construction of Social Reality* (New York: Free Press, 1995).
66. Jerrold Levinson, "What a Musical Work Is," *Journal of Philosophy* 77, no. 1 (1980): 28, https://jstor.org/stable/2025596.
67. Peter Lamarque, *Work & Object: Explorations in the Metaphysics of Art* (New York: Oxford University Press, 2010).

Appendix: Survey Questions

1. Semi-automated support may be useful in my professional tasks if I am able to shape the service through ongoing feedback.
   - Strongly agree
   - Somewhat agree
   - Neither agree nor disagree
   - Somewhat disagree
   - Strongly disagree
2. Which vocabularies would be most amenable to semi-automated suggestions for incorporation in your work? (Select all that apply)
   - LCSH
   - FAST
   - MESH
   - Other
3. If you could select a targeted language support for semi-automated subject suggestions, select those languages that would be most useful to your work. Select all that apply.
   - MARC language tag list https://www.loc.gov/marc/languages/language_code.html

4. Which genre/form areas would be most useful for semi-automated subject support?
   - History
   - Biography
   - Periodicals
   - Fiction
   - Congresses
   - Early works to 1800
   - Criticism, interpretation, etc.
   - Conference papers and proceedings
   - Maps
   - Sources
   - Exhibitions
   - Biographies
   - Bibliographies
   - Bibliography
   - Catalogs
   - Dictionaries
   - Poetry
   - Statistics
   - Sound recordings
   - Scores
   - Early works
   - Drama
   - Pictorial works
   - Electronic journals
   - Academic theses
   - Private bills
   - Video recordings
   - Internet videos
   - Handbooks, manuals, etc.
   - Others
5. If Annif could provide results based on a selected country of origin for a publication, please register which countries you would like to be able to select as a focus of semi-automated suggestions.
   - List of countries in the world (pre-loaded from Qualtrics country library)
6. Would you recommend the Annif service to a friend or colleague?
   - Yes
   - No
   - Maybe
7. What type of ongoing feedback would you be interested in participating in for shaping future Annif development?
   - In-person workshop or discussion
   - Virtual workshop or discussion
   - Not interested
   - Other