

Using Linked Data for Data Analytic Literary Research: Case BookSampo – Finnish Fiction Literature on the Semantic Web

Annastiina Ahola¹, Telma Peura^{1,2} and Eero Hyvönen^{1,2}

¹*Semantic Computing Research Group (SeCo), Aalto University, Finland*
<https://seco.cs.aalto.fi>, firstname.lastname@aalto.fi

²*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*

Abstract

The BOOKSAMPO Linked Data (LD) portal was deployed in 2011 by the Public Libraries of Finland and has today nearly 2 million annual users. Its LD covers virtually all Finnish fiction literature but the data has not been used for data analyses in Digital Humanities (DH). This paper discusses how the KG can be used for literary research in two ways: First, a new BOOKSAMPO 2.0 PORTAL user interface (UI) is presented, based on faceted semantic search with seamlessly integrated data-analytic tools for DH research as suggested in the Sampo Model. This application makes it possible to analyze the data without programming skills. Second, the BOOKSAMPO SPARQL endpoint API can be accessed directly by SPARQL querying and scripting, using tools such as Jupyter Notebooks. The analysis results presented suggest interesting spatial, temporal, and topical trends in how the Finnish fiction literature has evolved during the last decades. The approach and tools presented in this paper can be used for analyzing literary landscapes developments in other countries, too.


Keywords


Digital Libraries, Linked Data, User Interfaces, Data Analysis, Portals


1. Introduction

BOOKSAMPO – Finnish Fiction Literature on the Semantic Web¹ [1, 2, 3, 4] provides information on virtually all fiction literature published in Finland since mid-19th century. Its contents are based on rich semantic descriptions of books and their contexts using Linked Data (LD) that originates from library catalogs and related heterogeneous data sources. BOOKSAMPO data was originally part of CultureSampo [5, 6] (online since 2008) but has been maintained since 2011 independently by the Public Libraries of Finland (BLF). BookSampo has grown into one of the main web services of the BLF and is used by ca. 2 million users in a year.

BOOKSAMPO is an application instance of the more general “Sampo Model”² [7] for establishing LD services and creating semantic portals on top of them. There are over 20 Sampo portals in

 0009-0008-6369-4712 (A. Ahola); 0000-0003-1695-5840 (E. Hyvönen)

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

 CEUR Workshop Proceedings (CEUR-WS.org)

¹Portal: <http://kirjasampo.fi>; research homepage: <https://seco.cs.aalto.fi/applications/kirjasampo/>

²The model is called “Sampo” according to the Finnish epic Kalevala, where Sampo is a mythical machine giving riches and fortune to its holder, a kind of ancient metaphor of technology according to the most common interpretation of the concept.

use³ in Finland and beyond, based on a Linked Open Data (LOD) infrastructure for DH [8].

BOOKSAMPO is used by library users and librarians for finding literary works of interest and related contextual information. The original Drupal-based user interface (UI) in use since 2011 provides traditional text search engines for finding records and then related contents as links for data exploration. However, the full potential of the LD for searching, exploring, and for data analytic research has not been utilized in the original UI [3] and in later Sampo portals based on the Sampo-UI framework [9, 10]. Moreover, although the data service underlying the portals provides a SPARQL endpoint, this opportunity has hardly been used for analyzing the data in Digital Humanities (DH), as suggested in the Sampo model. This paper addresses these two issues as research questions:

1. How to develop and use a semantic UI, based on Linked Data, with seamlessly integrated data-analytic tools for DH research?
2. How to use literary Linked Data published in a SPARQL endpoint directly for DH research?

As a methodological solution approach, the Sampo model integrated with methods and tools of distant reading and Digital Humanities [11, 12, 13] are employed. The BOOKSAMPO knowledge graph is used as the case study data. As practical results of the work, the BOOKSAMPO has been published for everybody to use in a SPARQL endpoint by the open CC BY 4.0 license, and the new user interface on top of it is available on the Web, too. The Sampo-UI software used in our work has been published open source in GitHub.

This paper first introduces the Sampo model and Sampo-UI framework underlying BOOKSAMPO in Section 2, and presents the BOOKSAMPO KG in (Section 3). Section 4 explains the new semantic portal application on top of the KG with examples of using the system. Data analyses using the SPARQL endpoint are then presented and discussed in Section 5. After this related works are discussed (Section 7) and in conclusion (Section 8), contributions of the paper and case study are summarized, discussed, and next steps ahead are outlined.

2. Sampo Model: Publishing and Studying Linked Data

Table 1
Sampo Model Principles P1–P6

P1	Support collaborative data creation and publishing
P2	Use a shared open ontology infrastructure
P3	Make clear distinction between the LOD service and the user interface (UI)
P4	Provide multiple perspectives to the same data
P5	Standardize portal usage by a simple filter-analyze two-step cycle
P6	Support data analysis and knowledge discovery in addition to data exploration

The Sampo model⁴ [7] is a consolidated set of principles listed in Table 1 for collaborative publishing and using LOD on the Semantic Web. Principles P1–P3 lay a foundation for develop-

³See <https://seco.cs.aalto.fi/applications/sampo/> for Sampos, links, videos, and publications.

⁴The name “Sampo” comes from the Finnish epic Kalevala, where Sampo is a mythical machine giving riches and fortune to its holder, a kind of ancient metaphor of technology.

ing LOD services; principles P4–P6 are related to creating semantic portals. The model is based on the Semantic Web standards⁵ [14] and best practices of the W3C for Linked Data publishing [15, 16] and is supported by tools and infrastructures, such as the Sampo-UI framework for UI design. The model has evolved gradually in 2002–2023 when developing over twenty LOD services and portals that have had up to millions of end users, depending on the application⁶. Being domain-agnostic, the model has been used in a variety of application areas. For example: CultureSampo⁷ aggregates and publishes a wide variety of tangible and intangible cultural heritage collections; the HealthFinland system⁸ was used for publishing health promotion information; the Mapping Manuscript Migrations (MMM) system is an application⁹ for pre-modern manuscript studies; in BiographySampo¹⁰ and AcademySampo¹¹, biography and prosopography are in focus; NameSampo¹² is for toponomastic research on placenames; FindSampo¹³ is used for studying archaeological finds; LawSampo¹⁴ applies the model to publishing legislation and case law and ParliamentSampo¹⁵ is for studying parliamentary speeches, political culture, and networks of politicians.

From a LOD service development point of view (P1–P3), the model is based on the idea of collaborative content creation (P1)¹⁶. The data is aggregated from local data silos into a global service, based on a shared ontology and publishing infrastructure (P2). The shared ontology infrastructure includes 1) shared (*meta*)*data models* for representing data (e.g., Dublin Core or CIDOC CRM) and 2) a set of *domain ontologies*¹⁷ that are used for populating the instances of the data model classes, such as shared vocabularies for subject matter or historical places and actors. The local data are harmonized and enriched with each other by linking and reasoning. In this model everybody can arguably win, including the data publishers by enriched data and shared publishing infra, and the end users by richer global content and services. The model supports the idea of separating the underlying Linked Data service *completely* from the user interface via a SPARQL API (P3). This arguably simplifies the portal architecture and the data service can be opened for data analysis research in Digital Humanities. For example, the Yasgui¹⁸ [17] interface for SPARQL querying and visualizing the results can be used, or Python scripting in Google Colab¹⁹ and Jupyter notebooks²⁰.

⁵<https://www.w3.org/standards/semanticweb/>

⁶See <https://seco.cs.aalto.fi/applications/sampo/> for more info, publications, and videos about the Sampo portals.

⁷Portal online: <https://kulttuurisampo.fi>; project homepage: <https://seco.cs.aalto.fi/applications/kulttuurisampo/>

⁸This prototype was deployed in public use in Finland by the National Institute for Health and Welfare; project homepage: <https://seco.cs.aalto.fi/applications/terveysuomi/>

⁹Portal online: <https://mappingmanuscriptmigrations.org>; project homepage: <https://seco.cs.aalto.fi/projects/mmm/>

¹⁰Portal online: <https://biografiasampo.fi>; project homepage: <https://seco.cs.aalto.fi/projects/biografiasampo/en/>

¹¹Portal online: <https://akatemiasampo.fi>; project homepage: <https://seco.cs.aalto.fi/projects/yo-matrikkelit/>

¹²Portal online: <https://nimisampo.fi>; project homepage: <https://seco.cs.aalto.fi/projects/nimisampo/en/>

¹³Portal online: <https://loytosampo.fi>; project homepage: <https://seco.cs.aalto.fi/projects/sualt/>

¹⁴See LawSampo project homepage for more information and publications: <https://seco.cs.aalto.fi/projects/lakisampo/>

¹⁵See ParliamentSampo project homepage for more information and publications: <https://seco.cs.aalto.fi/projects/semparl/>

¹⁶In our case the collaborators are institutions rather than individual people.

¹⁷We use this term to refer to knowledge organization systems, such as SKOS vocabularies, that define terms and their relations in application domains.

¹⁸<https://yasgui.triplay.cc>

¹⁹<https://colab.research.google.com/notebooks/intro.ipynb>

²⁰<https://jupyter.org>

3. BookSampo Knowledge Graph and Data Service

The original user interface²¹ (UI) of BOOKSAMPO does not fully utilize the potential of the semantically rich underlying knowledge graph (KG), nearly 9 million triples today. The data covers virtually all Finnish fiction literature and beyond and is interesting from a Digital Humanities (DH) research perspective, too, for data analyses and visualizations. Table 2 lists the number of instances of different entity types in the data back in 2013 and ten years later in 2023.

Class (Type)	Instances	Class (Type)	Instances
Literary works	93,000	Literary works	215,000
Editions	127,000	Editions	222,000
Book covers	27,000	Book covers	119,000
Fictional characters	19,000	Fictional characters	49,000
Contemporary reviews	15,000	Contemporary reviews	15,000
Web links	10,000	Web links	25,000
Literary series	2,900	Literary series	8,900
Literary awards	2,700	Literary awards	6,400
Literary award series	200	Literary award series	300
Movies	1,100	Movies	2,000
People (e.g. authors)	29,000	People (e.g. authors)	64,000
Author's pictures	2,600	Author's pictures	4,200
Publishers	2,600	Publishers	5,500

Table 2
Instance counts in 2013 [1] (left) vs. 2023 (right)

Figure 1 illustrates how the novel *Pride and Prejudice* is modeled in the BOOKSAMPO KG. The `kaunokki:romaanii` entity represents the abstract work level of the novel. That entity has links to the entity of the author *Austen, Jane* and the `kaunokki:fyysinen_teos` entity representing the Finnish edition of the work. This Finnish edition further has the link to the publisher entity of the *WSOY* publishing house.

The BOOKSAMPO data divides works into two levels: abstract and physical work levels. The data model is based on the FRBRoo model [18] but simplified to the aforementioned two levels. The abstract work level is equivalent to the *work* level in the FRBRoo model while the physical work level represents what would be the *manifestation* level in the FRBRoo model. In practice the abstract work level deals with information that is shared between all editions of a work. The physical work level on the other hand contains edition-specific information, e.g., page number, and publisher, that are specific to the edition. In the case of the BOOKSAMPO data, this edition-specific information is often recorded for first editions in relevant languages. This means that an average work will have a physical work level entities for at least its first edition in the original language as well as entries for the possible first editions of translations of the work into Finnish and/or Swedish. Figure 2 illustrates the split for Mika Waltari's novel *Sinuhe Egyptiläinen*.

²¹<https://kirjasampo.fi>

PREFIX kaunokki: <http://www.yso.fi/onto/kaunokki#>
 PREFIX foaf: <http://xmins.com/foaf/0.1/>

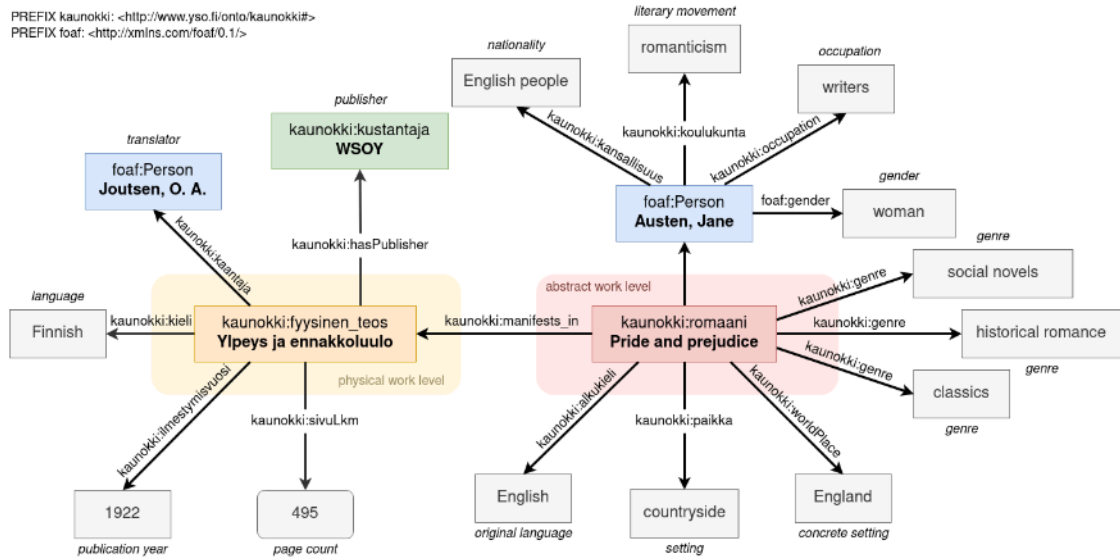


Figure 1: An example of how a novel is modeled in the BookSAMPO KG.

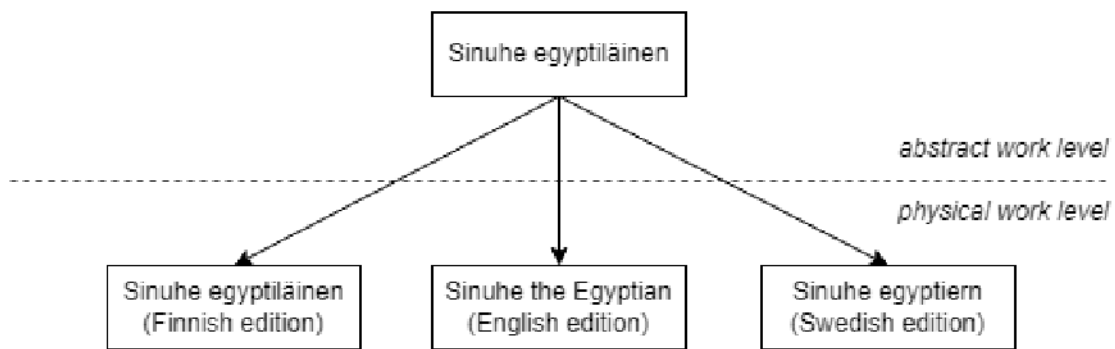


Figure 2: The abstract and physical work levels for Mika Waltari’s novel *Sinuhe Egyptiläinen*.

The BookSAMPO KG is available on the Linked Data Finland (LDF) platform [19], providing a home page for the KG, and a public SPARQL endpoint²². To support reuse, the home page provides additional information about the KG, such as, 1) schema documentation automatically generated by the platform, 2) sample SPARQL queries, and 3) metadata using *SPARQL Service Description*²³, and *Vocabulary of Interlinked Datasets (VoID)*²⁴. The LDF platform also provides dereferencing of URIs for both human users and machines, and a generic RDF browser for technical users, which opens when a URI is visited directly with a web browser.

²²The public SPARQL endpoint: <http://ldf.fi/dataset/kirjasampo/>

²³<https://www.w3.org/TR/sparql11-service-description/>

²⁴<https://www.w3.org/TR/void/>

The BOOKSAMPO SPARQL endpoint is hosted on an Apache Jena Fuseki²⁵ SPARQL server. The whole KG and Fuseki are contained in a Docker image, that can be easily built and started when and where needed.

4. New BOOKSAMPO 2.0 User Interface

This section describes the key ideas behind the BOOKSAMPO 2.0 UI design and shows how the UI is used for searching, browsing, and data analyses.

4.1. An Infrastructure for UI Design

The new BOOKSAMPO UI called BookSampo 2.0²⁶ [20, 3, 4] is built using the Sampo-UI framework. A key idea of Sampo-UI [9, 10], based on the Sampo model [7] (cf. Section 2), is to provide an easy to use framework for creating semantic portals on top of external SPARQL endpoints. The power of the tool comes from the possibility to re-use designs of earlier Sampo-portals in a new application, and include new components to the framework to be re-used by other projects later on. In a way, the framework aims at an infrastructure for portal design and at “standardizing” the UI model, which makes the framework easier for both the developers and end user to use. By learning how to create/use one Sampo, other Sampos can be created/used more easily due to the same UI structure and logic [4].

From a programming point of view, the configuration for a new portal is specified declaratively through JSON (JavaScript Object Notation) configuration files [10]. The Sampo-UI framework offers ready-to-use components to be re-used in semantic portals that can be added through the configuration files without the need for heavy coding. The used components can easily be expanded upon by adding new mapping functions and expanding upon configuration options passed to the components.

The Sampo-UI framework²⁷ is available in GitHub²⁸ under the open MIT License, and has been used successfully not only by the original developers but also by first external users. For example, the Norwegian place name service Norske stadnamn²⁹ is based on re-using the original Finnish NameSampo portal³⁰ [21].

4.2. Using BOOKSAMPO 2.0

The Sampo-UI framework facilitates the development of new semantic portals following the Sampo model’s principles by offering the data of a KG from multiple *application perspectives* using faceted semantic search and browsing combined with the filter-analyze two-step usage cycle. The user can search and browse the data by choosing a perspective and then filtering the data using the facet menus included in the framework. The resulting data can then be analyzed

²⁵<https://jena.apache.org/documentation/fuseki2/>

²⁶In use at: <https://analyysi.kirjasampo.fi/en/>

²⁷Project homepage: <https://seco.cs.aalto.fi/tools/sampo-ui/>

²⁸<https://github.com/SemanticComputing/sampo-ui>

²⁹<https://toponymi.spraksamlingane.no/nb/app>

³⁰NameSampo online: <https://nimisampo.fi>; project homepage: <https://seco.cs.aalto.fi/projects/nimisampo/>

by using the data-analytic tools (e.g., different statistics and visualizations) integrated in the Sampo-UI framework. In the original BOOKSAMPO 2.0 PORTAL only traditional text-based search with browsing is offered. In the new UI, text-based search is offered, too, but only as one facet among the others, making the Sampo-UI model more general.

When opening a Sampo portal, the user first lands on the *landing page* (shown in Figure 3) that provides access to the application perspectives. In our case there are five perspectives available:

1. *Novels*. This perspective deals with the abstract work level of novels.
2. *Publications*. This perspective deals with the physical work level of *all* works.
3. *People*. This perspective deals with authors and other people related to literature, e.g., illustrators, translators, and reviewers.
4. *Covers*. This perspective deals with book covers and information related to them.
5. *Nonfiction books*. This perspective deals with the abstract work level of nonfiction books.

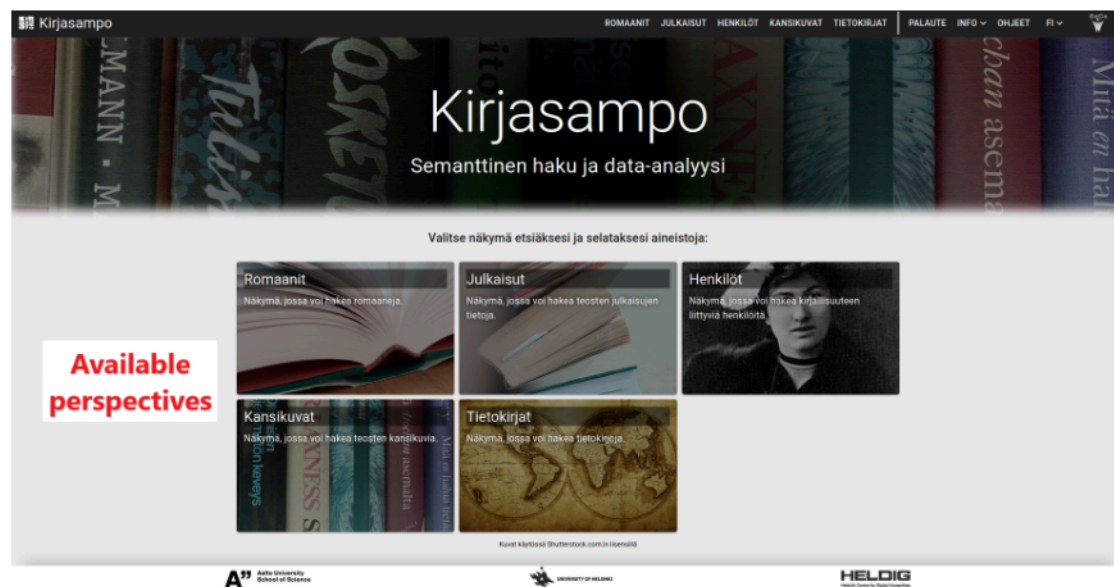


Figure 3: The landing page of the BookSampo User Interface.

All of the perspectives query the data from the same SPARQL endpoint of BOOKSAMPO KG. Each perspective is related to a class in the underlying data model and is used to search and analyse individuals of that class. For example, choosing the *Novels* perspective means that the class of novels (class `kaunokki:romaani` in Figure 1) are searched for. The faceted search view of a perspective lists properties of the class as facets. For example, in the case of novels there are facets for selecting authors, publishers, and topics of the novels. The search results are novel entities that match with the selections on the facets. After each selection on the facets a *hit count* is automatically computed for each possible next facet category selection. In this

way the user never ends up in a dead end of no found hits; this makes faceted search different from traditional search with filters.

The basis for choosing the five aforementioned perspectives was to cover aspects of Finnish literature as comprehensively as possible with few perspectives. *Novels* perspective was chosen as the perspective to cover fictional books due to novels being the largest subgroup of fictional literary works in the data. The *Nonfiction books* perspective was chosen to supplement the *Novels* perspective with nonfictional works, although in the data it only represents a non-comprehensive subset of nonfiction published in Finland. The *Publications* perspective covers all literary works on the physical work level. The split between (1) novels and nonfiction books and (2) publications follows the split made in the original BookSampo data [1] as introduced in Section 3.

To supplement the data on literary works, the *People* perspective was added to provide information on all people relevant to the presented data on literary works, whether it be the authors behind the books or other people relevant to them, such as the illustrators and translators of works. As the BOOKSAMPO KG also includes data on contemporary reviews, information of reviewers is also included in this perspective. To finish off the available perspectives, the *Covers* perspective was added due to the popularity of the book cover search function on the original BOOKSAMPO 2.0 PORTAL, the search capabilities of which could be even further improved with the use of faceted search.

Clicking on any of the cards for the different application perspectives on the landing page leads to that particular perspective's *faceted search view* (shown in Figure 4). The faceted search view consists of three important elements: (1) the facet menu on the left, (2) the results view on the right, and (3) the different visualization tabs on top of the results view.

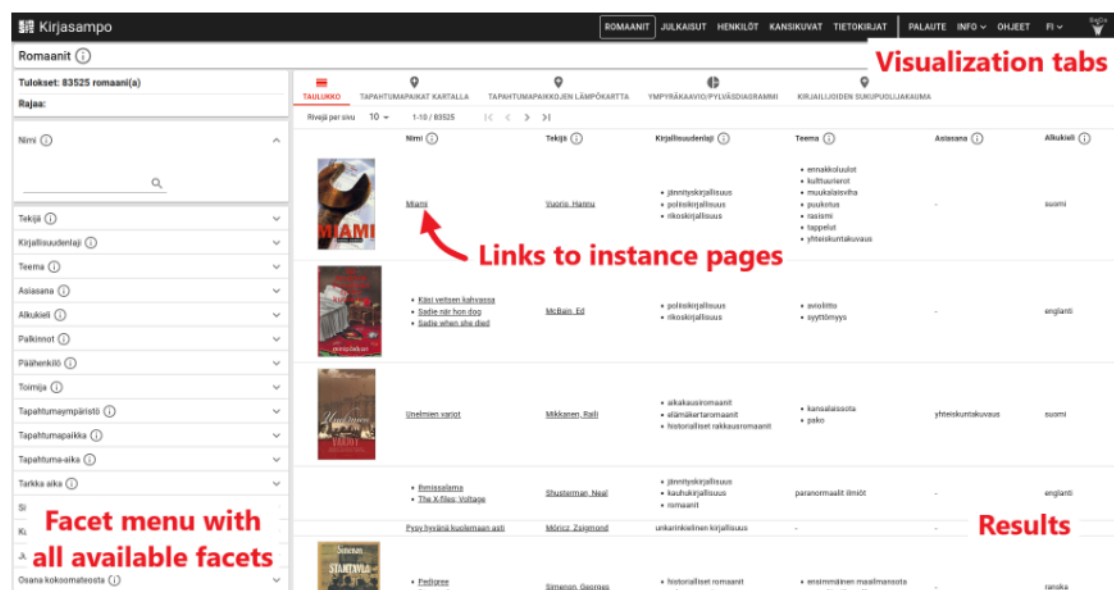


Figure 4: The faceted search view of the *Novels* perspective in the BookSampo User Interface.

The facet menu includes all the available facets that can be used to filter the data in that particular perspective. The Sampo-UI framework offers various types of facets for different types of data. The BOOKSAMPO 2.0 portal utilizes three different facets types:

1. *Checkbox facet*. A facet for filtering results by selecting one or multiple checkboxes for wanted property value entities. Results are automatically updated when a checkbox is checked. Selecting multiple checkboxes works in a disjunctive way: Selecting both the genres *romance* and *thriller* for the genre facet would return works that belong to either (or both) genre. If the property values have a hierarchical structure, e.g., yearly literary awards (*Finlandia Prize 2022*) and the award series they belong to (*Finlandia Prizes*), the facet can be configured to show the entities hierarchically, too.
2. *Integer range facet*. A facet for filtering results by limiting the integer range a property's value should be in, e.g., searching for works that have a *page count* in the range of 200–300. The facet is applied by pressing the 'apply' button.
3. *Text facet*. A facet for searching for results based on text string, e.g., searching for works by their *names*. The facet is applied after the user presses enter.

The results are shown in a table format on the right side of the screen by default. Each of the rows represents one entity. The different columns represent the different properties and property values these entities have. Column values that are underlined denote links to more information about that particular entity. By default the result set includes all entities that match the type of the application perspective with no filters applied. The results are automatically updated when any facets have been applied.

The links in the results table lead to the **instance pages** (shown in Figure 5) of entities. Instance pages aggregate all the information about that particular object in the same page. This includes the information shown about the entity in the table view as well as possible further information not deemed relevant to be included in the table view. In the BOOKSAMPO 2.0 PORTAL the general choice was to include information that could be used as facets as columns in the table view and leave the rest of the relevant information to the instance pages.

Similarly to the faceted search view, the instance page view can have multiple tabs for different ways of visualizing the data. These tabs depend on the type of the entity in question. Novels and nonfiction books, for example, have a specific tab for showing detailed information about the different publications of that particular work (shown in Figure 6) that exist in the data.

4.3. Integrated Data-analytic Tools

The various tabs available in the faceted search views and instance pages of application perspectives offer the user different ways of looking at the data as well as analyzing the data. The different visualization types available can be roughly split into three different categories:

1. *Pie/bar charts*. Charts that show the ratio of property values in comparison to each other, e.g., the top genres of novels (shown in Figure 7). The component behind these visualizations is created with the ApexCharts³¹ library.

³¹<https://apexcharts.com/>

The screenshot shows the Kirjasampo website interface. At the top, there is a navigation bar with categories like 'ROMAANI', 'JULKAISUT', 'HENKILOT', 'KANSIKUVAT', 'TIETOKIRJAT', 'PALAUTE', 'INFO', and 'OHJEET'. Below this, the page title is 'Romaani' and 'Miami'. There are two tabs: 'Table view tab' (active) and 'Publications tab'. The main content area shows the book cover for 'Miami' and a metadata table. A red text overlay on the right side of the image reads 'Information about the novel'.

URI	http://www.yso.fi/onto/kaurokkolates.50564
Nimi	Miami
Tekijä	Vuorio, Hannu
Kirjallisuudenlaji	kirjallisuus
Teema	ennakkoluulot
Aikajana	-
Aikaväli	suomi
Palkinnot	-
Päähenkilö	Parviainen, Markus
Toimija	jenpjt
Tapahtumaympäristö	ajamaailma
Tapahtumapaikka	Helsinki
Tapahtumajono	-
Tarkka aika	2001
ISBN	951-0-28381-9

Kaikki saa alkunsa siitä, kun kolmannelle luokalle kuuluu käynä Omar tömää rasistiseen nahkatakkimallikoon. Vaikka Omar seivää välikohdauksesta naarmulta, hänen vanhemmat veijensä Ali ja Abdi päättävät kostaa pelotellun nahkatakkelle. Pian yksi Omaria pelotellusta nahkatakkesta löydetään kuolleena, ja keston kierre alkaa nahkatakkien ja somalien välillä Malmilla. Samalla Malmilla rikospoliisi selvittää niin lasten pelotelta kuin kuolemantapauksien. Tutkimuksessa on jännäisiä löydöksiä, koska sitä selvittävät kokemieet poliisit Väara ja Parviainen kilpailevat ylennyksestä, ja tapauksen ratkaiseminen voisi kallistaa vaakakupin toisen puolelle. Vaan osako poliisikaan ennakoita, mitä kuolemantapauksesta voi seurata? Hannu Vuorion Miami on Helsinkiin sijoitettu ohjelma, jossa kuolemantapauksista ja somalien törmäyksistä.

Figure 5: The instance page of Mr. Hannu Vuorio’s novel *Miami* in the BookSampo User Interface.

2. *Maps*. Charts that show entities on a map based on some location information related to the entity, e.g., settings of novels on a map (shown in Figure 8). The components behind map visualizations are created with the Leaflet³² and deck.gl³³ libraries.
3. *Time series*. Charts that show the evolution of entities or some of their properties as a function of time, e.g., the evolution of average page counts throughout years (shown in Figure 9). The components behind these visualizations are created with the ApexCharts and AmCharts³⁴ libraries.

4.4. Example Use Cases

This subsection presents two example use cases for the new UI. The first one focuses on finding a singular entity of interest. The second example showcases how the UI can be used to explore the general trends in the Finnish literature.

4.4.1. Example Use Case: Finding Works Fulfilling Specific Criteria

Figure 10 showcases an example use case where a user is looking for a novel fulfilling specific criteria. The illustrated steps of the usage cycle are the following:

1. The user selects the Novels perspective to search and browse novels.
2. The user then makes the following choices in the facet menu:

³²<https://leafletjs.com/>

³³<https://deck.gl/>

³⁴<https://www.amcharts.com/>

The screenshot shows the Kirjasampo website interface. At the top, there is a navigation bar with categories like ROMAAINIT, JULKAISUT, HENKILÖT, KANSIKUVAT, TIETOKIRJAT, PALAUTE, INFO, OHJEET, and FI. Below this, the search results for 'Miami' are shown. The book cover for 'Miami' is displayed, along with a table of metadata:

Nimi	Miami
Kustantaja	WSOY
Julkaisuvuosi	2003
Sivumäärä	307
Kieli	suomi
Ensimmäinen versio	kyllä
Muut tekijät	-

To the right of the table, there is a red text overlay that reads: "Information about the publications of this particular novel".

Figure 6: The list of editions of Vuorio Hannu’s novel *Miami* in the BookSampo User Interface showing the singular Finnish edition of the work.

- a) The *genre* of the novel should be *romance*.
- b) The *characters* in the novel should be *nobility*.
- c) The *setting* should be *castles*.

The results and hit counts are automatically updated each time the user makes a selection in the facets for the above criteria.

3. The user looks at the list of results and finds one that looks interesting. The user then clicks on the name of that novel.
4. The instance page of that particular novel is opened. Here the user can see more information on that particular novel and choose whether or not this novel is something (s)he might want to, for example, read. If this particular novel is not a match, the user can just go back to the search and look at other novels fulfilling the specified criteria.

4.4.2. Example Use Case: Exploring the Evolution of Finnish Literature

Figure 11 illustrates another example use case where the user wants to see how the popularity of the top 10 themes and keywords used for Finnish novels has evolved throughout the years. The illustrated steps of the usage cycle are the following:

1. The user selects the Publications perspective to browse publications.
2. The user selects the wanted facet values:
 - a) The *language* of the publication should be *Finnish*.
 - b) That *version* of the publication should be the *first (original) version*.
 - c) The *type* of the work should be a *novel*.

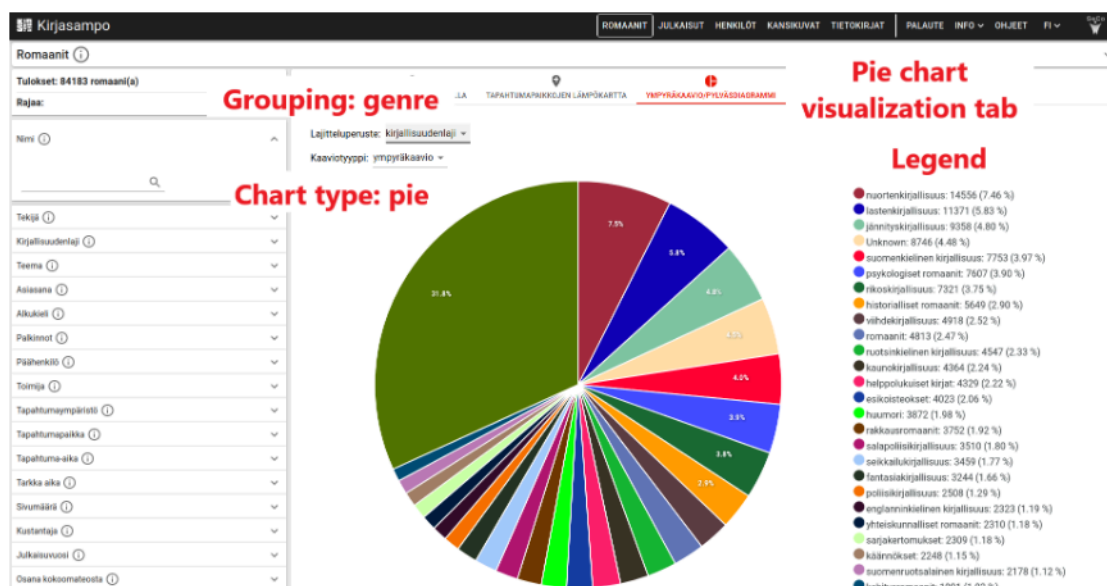


Figure 7: The top genres of novels visualized as a pie chart.

The results will again automatically update for the user after the facet selections. The results now show only the first publications of novels that were originally written in Finnish, i.e., excluding translated works.

3. The user then selects the visualization tab for the annual themes and keywords visualization. The user can there further zoom in and out on different time periods as well as pan the timeline.

4.5. Evaluation

Previous use case examples illustrate the opportunities and challenges of the Sampo-UI model for DH research. The usability of the demonstrator UI has not been evaluated formally by external users. However, Burrows et al. [22] report on evaluating a similar kind of Sampo-UI-based UI using the Mapping Manuscript Migrations (MMM) portal [23] where pre-modern manuscript collections are used as data. As the UI logic in all Sampo-UI systems is similar [10], the results suggested promising usability of the Sampo model and its UI logic from an end user's point of view, but of course the MMM portal is different from the BOOKSAMPO demonstrator. An empirical indication of the usability of the Sampo systems is that they have now been used in several different online CH portals that have had over million users on the Web in total³⁵.

Evaluations regarding using faceted semantic search and browsing, the basis of the Sampo UI model, suggest that this search paradigm is very usable when the user does not know exactly what (s)he is looking for [24, 25]. Otherwise, traditional string based searching is usually

³⁵Information about the Sampo portal series is available at: <https://seco.cs.aalto.fi/applications/sampo/>.

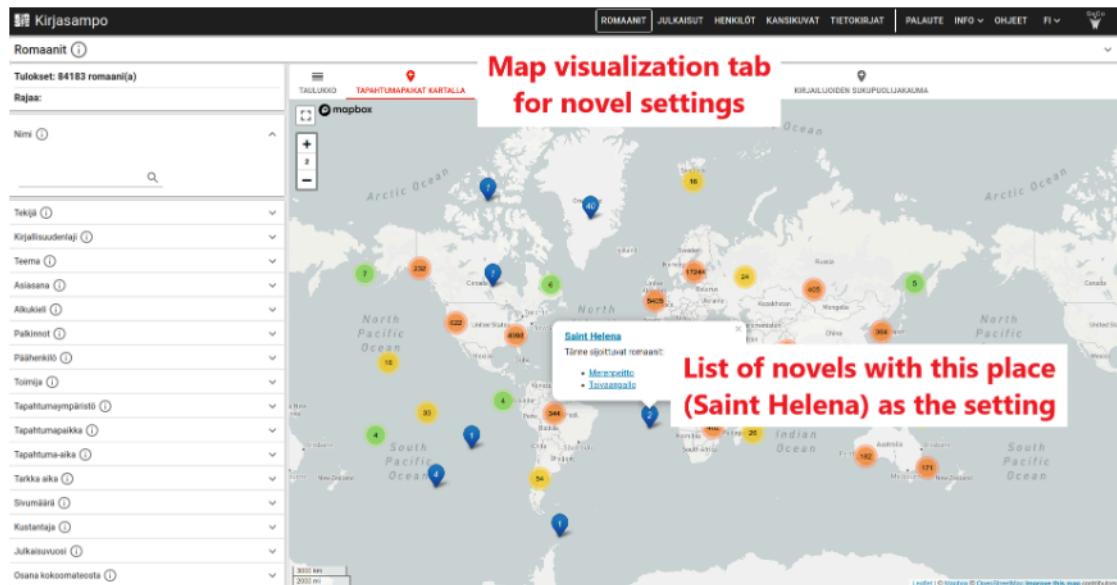


Figure 8: The settings of novels on a map.

preferred. To cater both needs at the same time, Sampo-UI has specific text search functionalities available, i.e., both search paradigms can be supported.

As for computational complexity, the Sampo-UI tool has been shown to scale up to hundreds of thousands of search objects (class instances) but the complexity depends on the data model used and how many and how large hierarchical facets are used [9]. A computationally demanding task in faceted search is pre-computing the hit counts for each facet category after each filtering step. When dealing with very large instance sets it is possible to force the user to constrain the search space first and use faceted search only after that. In NameSampo, for example, over two million placenames related to places are considered, and the search focus is initially constrained by limiting the area in question on a map or by text search on names [21].

5. Case Study – Literary Diversity through Metadata

To explore the potential of the BOOKSAMPO data in DH research, the KG was used to analyze the diversity of novels published in Finnish in the last 50 years [26]. The analysis was limited to publications annotated as novels published in Finnish during the period of 1970–2020, encompassing works written originally in Finnish as well as those translated to Finnish during that time period. The aim was to analyze diversity in terms of extra-textual indices annotated in the metadata: translated languages, author nationalities, author gender, and novel genres.

The theoretical framework of the data analysis was based on the following concepts:

1. *Distant Reading* [11] to adopt a more quantitative perspective to literature and to capture large patterns in the literary production. Instead of analyzing the text contents, the focus

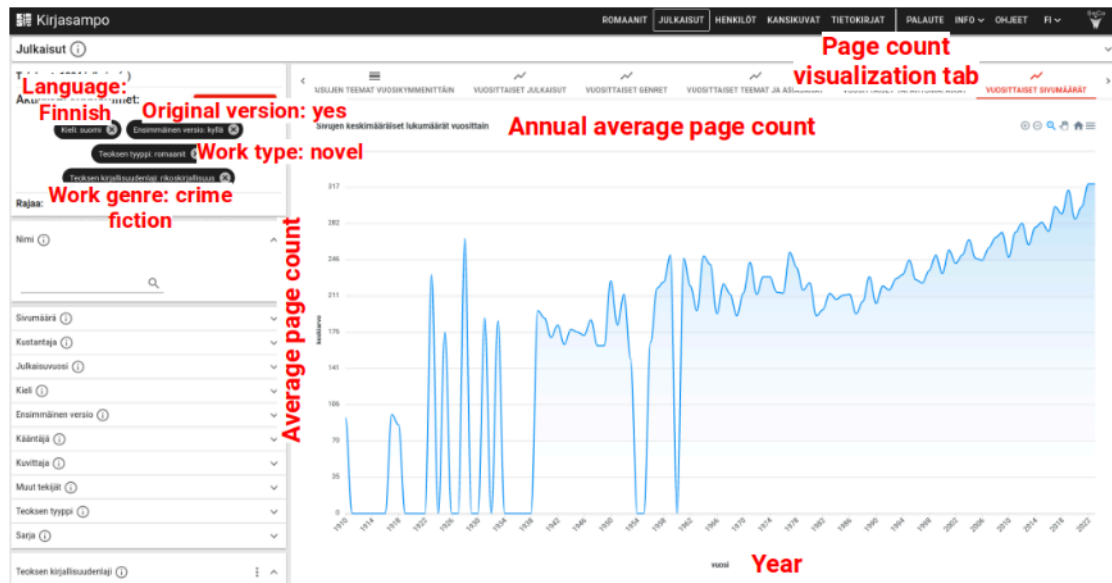


Figure 9: The evolution of average page count of publications.

was on extra-textual indices and the broad points that could be inferred from the mass of data.

2. *World Literary Space* [27] to adopt the metaphors of center and periphery in the Finnish literary space: which cultures and languages dominate the center (i.e., have more cultural capital) and what is regarded as peripheral, having less value?
3. *Translations as Cultural Transfers* [28] to study the effects of globalization on Finnish literature: is globalization acting as a mean of opening up diversity or is Finnish literature becoming even more dominated by mass cultures instead?
4. *Transnational, Entangled Literature* [29] to study how tightly the idea of 'Finnish literature' is tied to the Finnish language and nationality in the globalizing world.

In brief, the main goals in the current case study was to form an overall view of the trends in the diversification of the published novels, through the following questions: If we look at the Finnish literary space, what kind of centers and peripheries are there? How have they developed in the past 50 years? What languages and nationalities dominate the space according to the publication metadata?

To answer these questions, the data was analyzed from three perspectives: (1) the balance of novels originally written in Finnish and those translated into Finnish, (2) the diversity of published author nationalities and translated languages, and (3) diversity through genre hybrids and new genres.

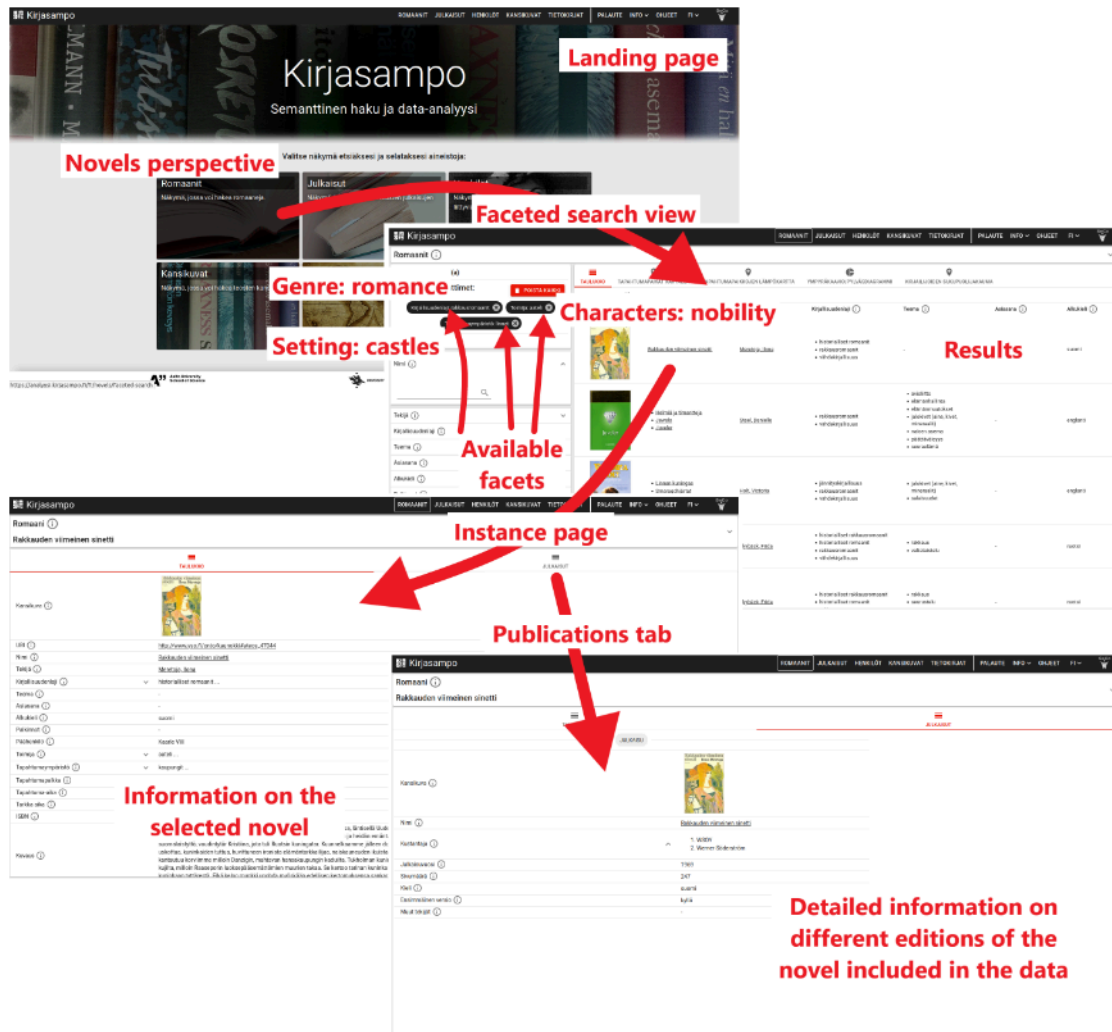


Figure 10: An example use case of the BookSAMPO 2.0 PORTAL where the user wants to find a novel that matches a set of criteria.

5.1. Compiling a Dataset

As extra-textual markers, the following annotations were queried from the SPARQL endpoint for all novels published in Finnish:

- Original language
- Publication year
- Author gender
- Author nationality
- Genre

The query results were preprocessed before performing the analyses to account for some of

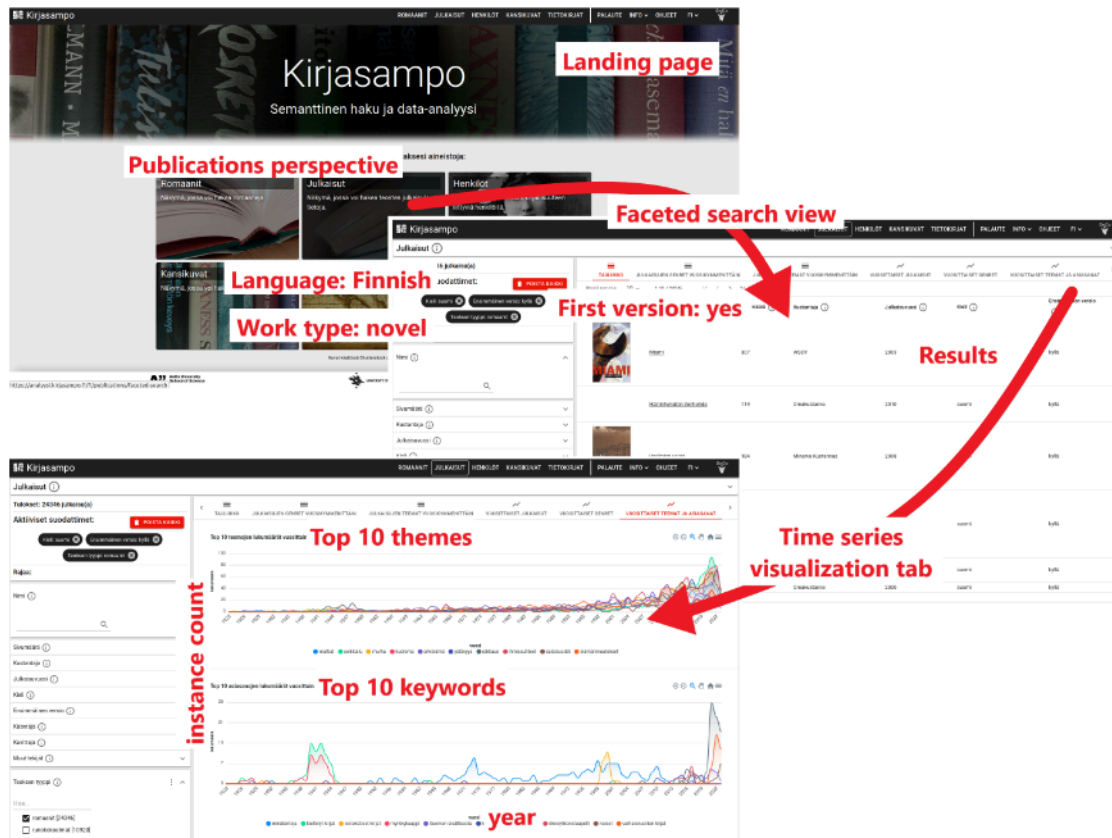


Figure 11: An example use case of the BookSampo User Interface where the user wants to explore how the most popular themes and keywords have evolved for Finnish novels throughout the years.

the peculiarities in the data³⁶. For instance, multiple entities for one nationality or national identities that for the purposes of this analysis could just be aggregated were combined under a shared label (e.g., 'Scottish' and 'English' people under the 'British' nationality). The consistency and completeness of annotations over time was checked prior to carrying out the analyses.

The final number of works and their annotated properties are listed in Table 3, showing that most publications were annotated for the selected diversity markers. The evolution of the number of publications in the investigated time period is illustrated in Figure 12.

5.2. The Status of Finnish as a Literary Language

In terms of published titles, the Finnish literary space seems to have become more domestic. As Figure 12 illustrates, the numbers of published Finnish works and authors has surpassed the respective numbers of translated works and authors since the 2010s that had been dominant before. This indicates an interest and trust towards Finnish literary culture, strengthening the value of Finnish as a literary language.

³⁶All code can be found on <https://github.com/telmauu/ks-tutkielma>

Type	All	Nationality known	Gender known	Publisher known	Popular fiction
Finnish novels	16,506	16,441	16,305	16,210	4,174
Translated novels	17,317	16,460	16,936	17,189	8,207

Table 3

Number of instances and the number of instances annotated with a specified property for the filtered data set. The last column shows the number of publications annotated as popular fiction, analyzed in subsection 5.3

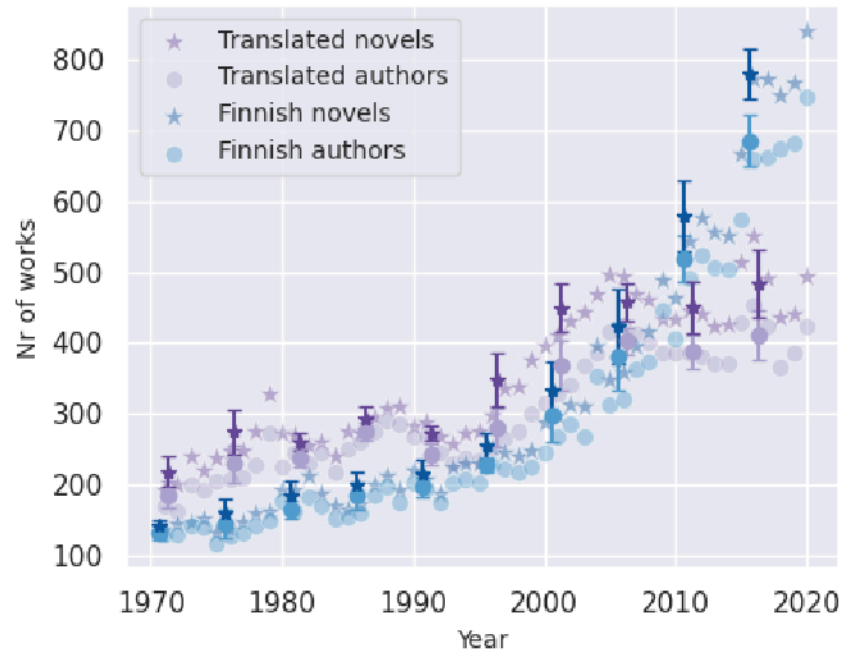


Figure 12: Evolution of the number of publications and authors. 5-year means with standard deviation are plotted in darker shades.

However, when dividing the published novels into several publisher types by size, the results also point towards a change in the underlying publisher field. Figure 13 depicts the evolution of publication numbers grouped by the type of the publisher (traditional publishers separated by size, self-publishing and other), and shows that although the proportion of novels having Finnish as the original language has grown in all categories, the most striking difference stems from the increase in the *self-published* and *other* categories.

In addition to the *self-published* and *other* categories, the number of works published by small-sized traditional publishing houses has increased for both Finnish and translated novels to approximately the same level with the large-sized publishers. This indicates a shift in the power balance in the Finnish literature publishing scene, where the traditional 'gatekeepers' of

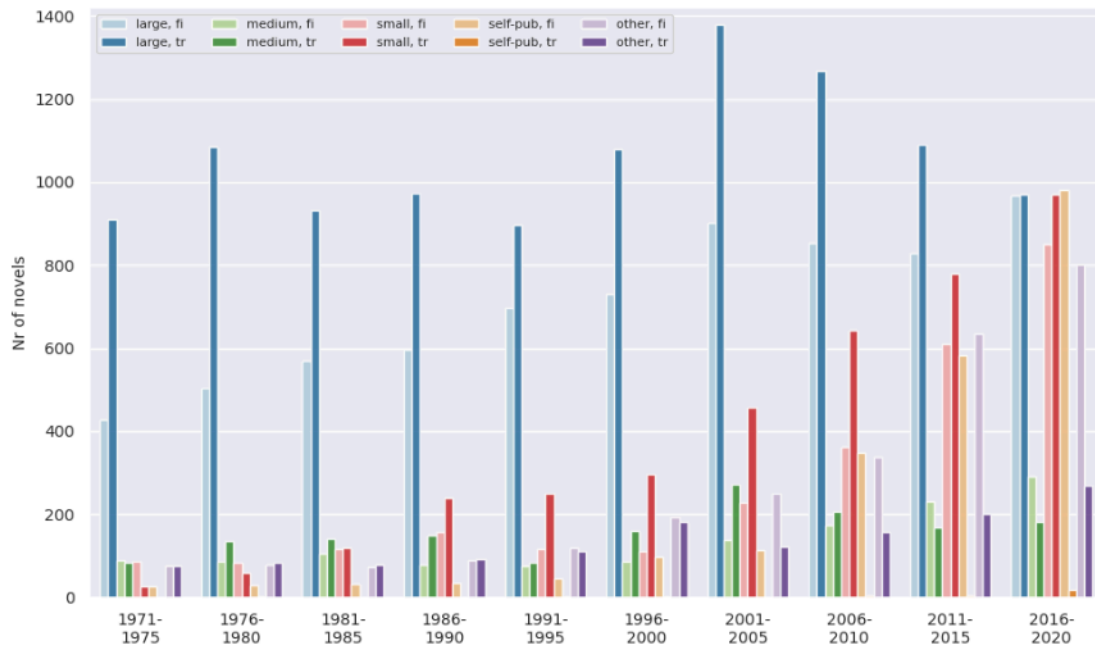


Figure 13: The number of published novels in 5-year periods, divided by publisher type and original language (Finnish/translated).

bigger publishing houses no longer wholly dominate the scene. Rather, there are new ways for people and different voices to get published through smaller publishers and self-publishing.

5.3. Literary Diversity of through Languages and Authors

The binary division into the groups of Finnish/translated language above was necessary to consider how the Finnish literary space was distributed between different author nationalities and original languages. The parallel analysis of author nationalities and languages indicates that the Finnish Literary Space is still centered around a few dominant linguistic and cultural groups.

First, the findings suggest that novels originally written in Finnish were mostly written by native Finnish-speaking Finns. Figure 14 illustrates the appearance of national and linguistic background of authors writing in Finnish. Though the number of nationalities³⁷ has been increasing throughout the examined time period, many of the non-Finnish nationalities annotated for writers are from areas geographically close to Finland even up to around 2010s. Only in the last investigated decade, the results showed indices of authors becoming more transnational and culturally entangled.

For translated works, both the original language as well as the nationalities of the authors were relevant to the diversity of Finnish literature. Moreover, the author gender was considered

³⁷It is important to note that linguistic groups do not correspond to nationalities. For example, Swedish- or Sámi-speaking Finns are still included as their own groups.



Figure 14: Nationality background (other than Finnish) of authors writing in Finnish. A bigger circle indicates a higher number of authors with that particular nationality annotated.

to get an idea about whether the diversity differed between the two genders. The absolute number of different languages and nationality backgrounds has increased. However, when looking at their relative shares, the picture looks slightly different, as Figure 15 illustrates.

While the relative proportion of Anglo-American literature has started to decrease after the 1990s, it still dominates the field of translated literature. The analysis shows that the proportion of 'other languages' (colored gray, for languages with 100 or less published works in the data) has stayed relatively constant throughout the years. Moreover, the proportion of already-established literary languages cultures, such as the Nordic neighbours, account for the decreasing proportion of Anglo-American, instead of the giving space to 'new' languages or cultures in translation literature. As to nationalities, the proportion of 'other' nationalities seems to even have decreased. The findings also indicate that language borders largely follow nationality borders without notable entanglement.

As to author genders, while the diversity of languages and nationalities has been increasing for both male and female authors in the examined time period, the situation is still biased, favoring male authors. The left diagram in Figure 16 illustrates the absolute number of languages and nationalities that have been annotated for translated works in the specified time period divided by gender. The diagram on the right of the figure illustrates the Simpson diversity index [30], taking into account both the total number and the share of different languages/nationalities in the data that was used for the same variables. These diagrams highlight a statistically significant systematic bias towards the cultural diversity of male authors over female authors in

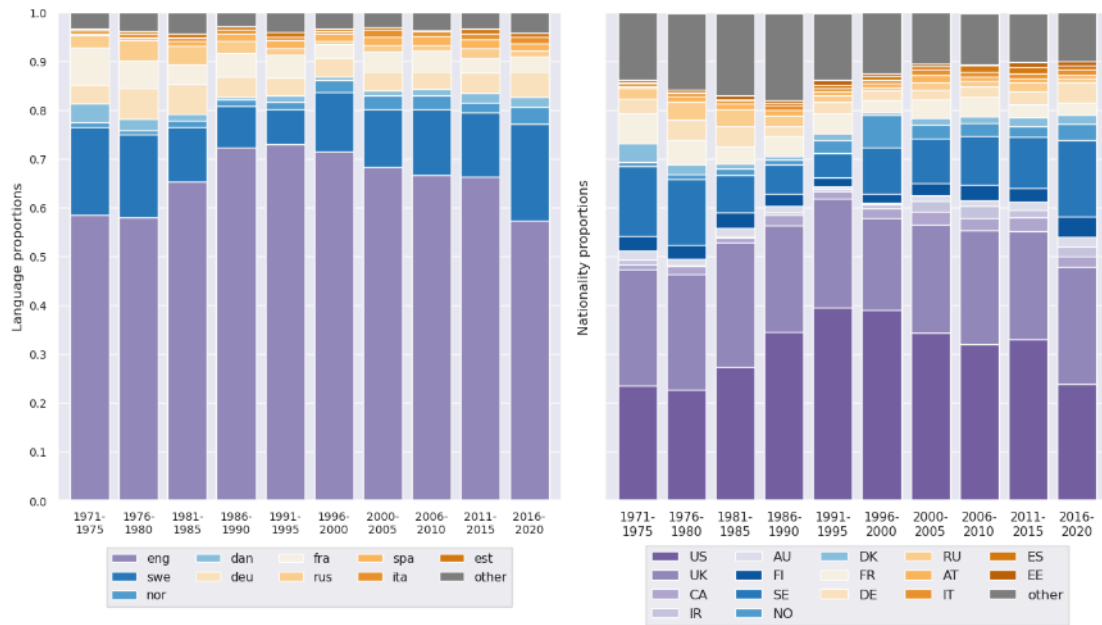


Figure 15: The relative evolution of languages (left) and nationalities (right) for authors of translated works.

the examined fifty-year time period.

5.4. The Effect of Genre

Another important point that was observed during the analysis was that not all genres are created equal—the diversity of languages and nationalities differs between genres. So, instead of there being a single center and periphery as in Casanova’s Literary Space model [27], our findings suggest that the Finnish literature space can be divided into multiple subcenters and peripheries. For example, crime literature has developed into a strong Nordic center from 10% in the 1990s to 40% in the 2010s, whereas Anglo-American female authors dominate the romance genre at around 80%.

Moreover, the absolute number of languages and nationalities for novels annotated as popular fiction is considerably smaller compared to the overall result. Even if we look at the genre that seems the most diverse, suspense fiction, in reality it is still very much dominated by two cultural groups: Anglo-American and Nordic. Translations from the Anglo-American and Nordic sphere account for around 90% of the translated works from the 2010s. With Finnish works also making up around a half of the total suspense novels, the share of non-Anglo-American and non-Nordic works becomes minimal. The rise of Finnish suspense literature in the BOOKSAMPO data is in line with the observations that suspense literature has been growing rapidly in 2000s [31] as well as with the international rise of popularity of Nordic suspense literature [32], especially that of *Nordic Noir* [33, 34].

In addition, it is important to keep in mind that genre definitions evolve over time, and what

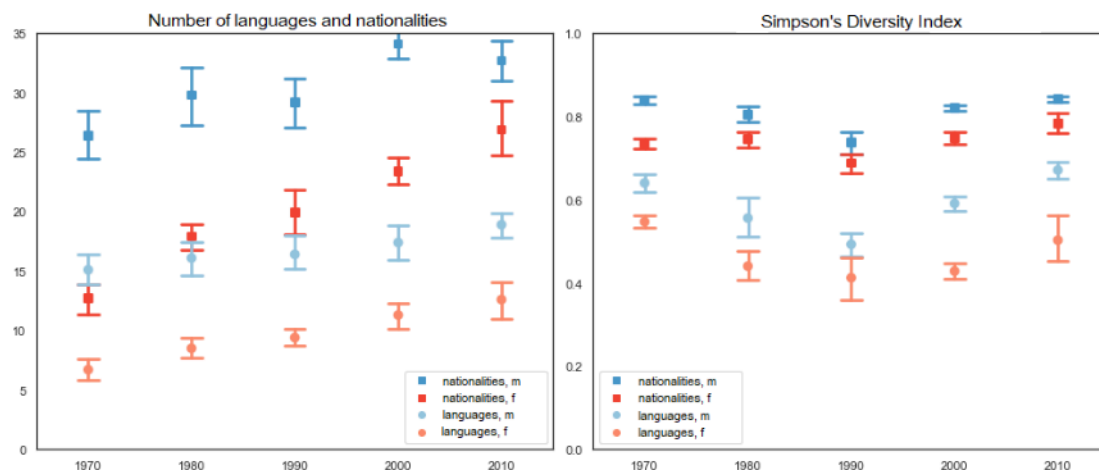


Figure 16: Number of different languages and nationalities divided by gender throughout decades and their Simpson diversity index for translated works.

has been annotated to a certain genre throughout the examined time period might have changed, skewing the numbers. The next section addresses this point more closely.

5.5. The Arrival of Genre Hybrids and New Genres

The BOOKSAMPO data can be also used to illustrate the arrival of new genres to the Finnish literary space. Figure 17 illustrates the evolution of number of publications for translated (darker shades) and Finnish novels (lighter shades) of the genres *fantasy*, *science fiction*, *dystopia*, *new weird* and *speculative fiction*.

From the graphs we can see that, initially, translated works clearly surpass the numbers of Finnish novels for the genres of interest. This observation fits with Moretti's model of literary waves [35]: new genres arrive in a national literature space through translations—the genre first gains a definition through the translated works before being adopted and adapted in the domestic scene. For example, in Figure 17 the fantasy genre is first annotated for some classic fantasy works like Tolkien's *Lord of the Rings*, Lewis' *Narnia* and Le Guin's *Earthsea* and then spikes again in the late 1990s after the first works of *Harry Potter* are translated into Finnish, which lead to Finnish publishers looking to publish more fantasy novels for children and teens [36]. Eventually, for both science fiction and fantasy, the number of Finnish works caught up to the translated works, or even surpassed it as in the case of fantasy literature, where in the latter half of 2010s there were a higher number of Finnish fantasy novels than there were translated ones.

The adaptation of genres to the local market is another interesting aspect of the emerge and evolution of genres. For example, the *dystopia* genre has gained significant fraction in the Finnish literature space in the last years. Moretti's [11] metaphor of trees could be applied here for the local evolution of this genre: the genre has absorbed some local elements in a periphery, the Finnish cultural context, and the domestic dystopia novels have certain more unique aspects

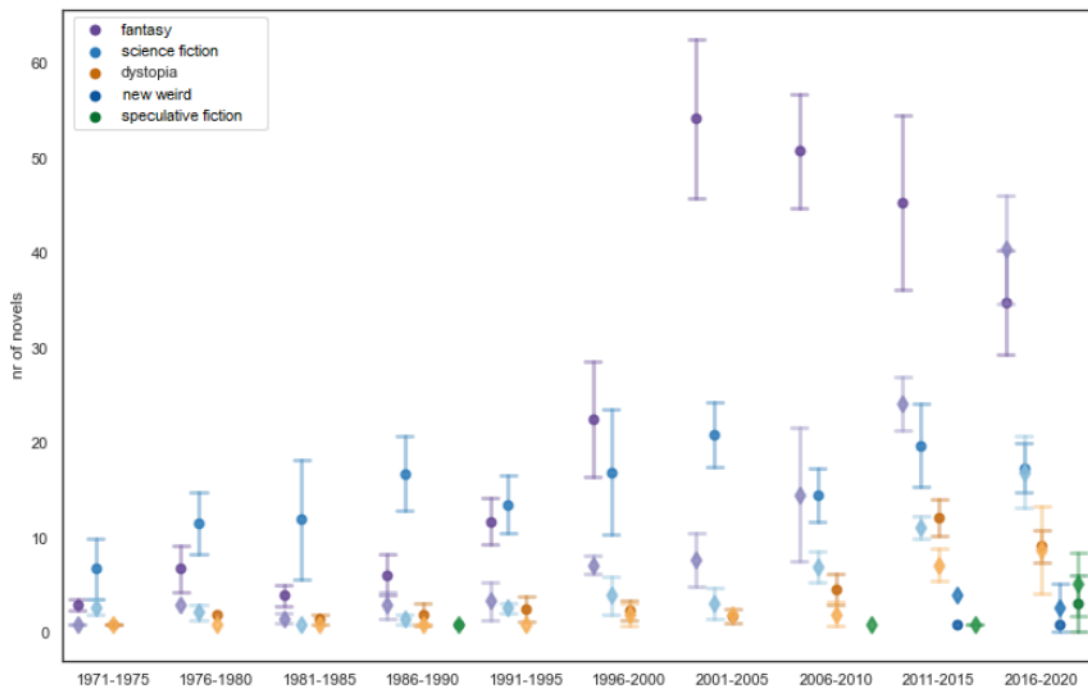


Figure 17: Evolution of number of publications for translated and Finnish novels for specified genres with darker shade of a color representing the translated works and a lighter shade the Finnish works

(e.g., a more northern location as a setting as well as increased importance on nature in the works), manifesting in new local subgenres of dystopia: *climate dystopia* and *ecodystopia*.

However, in the context of BOOKSAMPO data, there are no annotations for these two genres, even though they are discussed in literature research in Finland [37]. This likely speaks to both the delay in the definition of new genres as well as an additional delay until the genre is accurately identified in the annotation. The terms *climate dystopia* and *ecodystopia* are relatively new in an academic setting [37], so novels fitting these genres like couldn't be identified to belong to these particular subgenres and were instead annotated with the supergenre *dystopia*, while future similar works might get annotated with the more specific subgenres.

Even if annotation terms for particular more niche genres exist in the BOOKSAMPO data, the classification is still up to the person doing the annotations and even similar works can end up with different annotated genres. For example, *Teemestarin kirja* (2012) and *Kudottujen kujien kaupunki* (2015) by the same author, Emmi Itäranta, are annotated into different genres: *dystopia* and *science fiction* for *Teemestarin kirja* and *fantasy* and *new weird* for *Kudottujen kujien kaupunki* respectively. The genre *speculative fiction* in general is sparsely annotated, even though the term has been in use for decades at this point [38]. Something like the classification of genres will always be affected by the person annotating works due to the subjective aspects of exact genre determination.

6. Discussion

The new UI has been available to the public since October 2023, after being in an internal development phase, where it was tested by the members of the Semantic Computing Research Group as well as by people of the Finnish Public Libraries. The idea of BOOKSAMPO 2.0 PORTAL is to provide an alternative data-analytic UI for the BookSampo KG, complementing, not replacing, the original versatile legacy application with novel features. During the testing phase, the most important data issues of the underlying KG were corrected as they were identified by the new UI that makes the data more explicit to the user (e.g., missing labels and language tags, multiple instances of the same entities, etc.). Also the UI was improved and modified based on the feedback and discussion between the research group and the Finnish Public Libraries, but the UI has not been formally evaluated. However, based on previous implementations of semantic portals utilizing the Sampo-UI framework, the framework is however suggested to have good usability and scalability for the end user [22, 9, 39]. From a software developer point of view, the Sampo-UI framework was deemed very useful by the main developer of the portal who was not involved in developing the tool [20] before.

During this work, the new corrected KG of BOOKSAMPO has been made openly available using the CC BY 4.0 license, and was published on the Linked Data Finland platform³⁸ [19]. This data publication makes it possible for DH researchers to make more analyses like the ones presented in Section 5, and for software developers to create applications, such as the BOOKSAMPO 2.0 PORTAL presented in Section 4.

The Sampo-UI framework's structure enables easily developing future improvements by offering various ways to extend and customize the portal through features like custom components based on the needs of the users. This flexibility was tested and demonstrated already during the development work based on the comments from the members of the research group as well as the people from the Public Libraries. Possible problems or lacking features could be implemented and/or fixed in the future when they are identified. After the portal was made public, some tweaks and improvements in the functionality have also been made based on user feedback received by the Finnish Public Libraries.

Based on the usage statistics, the new BOOKSAMPO 2.0 PORTAL has around a dozen users daily. The implemented analytics are limited in what they track due to privacy reasons, but based on the feedback received, users are formulating their own queries and are looking to share these with other people. This wish for a sharing functionality was noted by the Finnish Public Libraries and the wish was recorder to be developed into a new feature in the UI in the future.

The case study on diversity, performed on the BOOKSAMPO data gave indications of Finnish literature in general having become more diverse in the last five decades. Overall, the results are encouraging in regards to the future of Finnish literature: the publication numbers—especially of works originally written in Finnish—are rising during the whole 50-year period. This finding contradicts previous fears of Anglo-American literature becoming even more dominating in Finland [40]. Nevertheless, the strong position of Finnish language could also be seen as an indication of domestic literature not really reflecting the reality of today's multicultural Finland, but rather representing only people who write in the official languages of Finnish or

³⁸BookSampo data homepage: <https://www.ldf.fi/dataset/kirjasampo/>

Swedish [29], suppressing cultural minorities.

As to translated works, Finnish literary space seems to become more diverse in terms of author backgrounds and translated languages. This diversity, however, is biased by genre. When looking into diversity within popular fiction, the works being translated were largely centered around dominating Anglo-American and Nordic groups. This could be due to popular fiction being viewed as something to consume for entertainment—not something that should challenge the reader or having high 'literary capital' or prestige. Following this line of thought, the market value might become more important than the literary value when making translation decisions. From that point of view, it is positive that the publisher field is growing in the number of different actors, as this might foster diversity through competition in the future.

The data analysis performed in the scope of this project only focused on the metadata, so textual analysis on the content of the books present in the data would likely give more insights on the changes that have happened in the Finnish literature scene in the last five decades in terms of diversity. The quantitative nature of the analysis done here also required some compromises made with the choices of, e.g., using the officially annotated nationality, binary choice of gender³⁹, as well as original language as measures of diversity as they might not paint the whole picture of the diversity of the author identities or the contents in the novels.

In addition to the annotations not necessarily giving all the relevant information about a person or a work, it also should be noted how subjective or difficult some aspects might be to annotate. New genres are born and defined all the time, and people might appreciate different qualities in a text, which will reflect on the annotations of different annotators. The existence of the hand-crafted annotations on things like genre in the first place do provide added value to the data and with more development and unification could potentially be used in future research to look at Finnish literature in a more general sense instead of focusing just on novels from a certain time period. The performed analysis acts as proof that this data set could be used for Finnish literature research, which it previously hasn't been properly used in. The data set has been used in Hackathons to answer some general questions about Finnish literature and its evolution, but never to this extent as in this project. In a way, this data analysis could be seen as a 'gambit' to discuss the evolution of Finnish literature by illustrating the broad lines of what has been happening in the bigger picture of the Finnish literature scene.

7. Related Works

Linked Data and ontologies have been used in libraries [41], museums, and archives [42, 43]. Using LD is advocated by major library organizations, such as IFLA⁴⁰ and OCLC⁴¹, and several libraries provide their collections as data in this form [44]. LD has been used in building

³⁹The binary division of 'male' and 'female' in the analysis is a limitation of the BOOKSAMPO data. There are no currently annotation options for non-binary gender identities in the data and people with uncertain gender identities or those falling outside of the gender-binary have been left with an unannotated gender.

⁴⁰<https://www.ifla.org/references/best-practice-for-national-bibliographic-agencies-in-a-digital-age/service-delivery/linked-open-data/>

⁴¹<https://www.oclc.org/research/areas/data-science/linkddata/linked-data-overview.html>

infrastructures, such as ARIADNEplus⁴² for archaeology, Linked Art⁴³ for fine arts in the U.S. and beyond, and in local efforts in Italy [45], the U.K. [46], and Finland [47, 8] to list a few examples. Cultural Heritage and DH have become a major application domain for LD technologies [48]. However, there has been less research on how to use the LD through intelligent UIs [7].

The potential of metadata has also been recognized in digital literary studies. The analysis and interpretation of bibliographic metadata allows for studying long-term patterns in literary history [49]. This quantitative approach fits well with the idea of distant reading, the application of computational methods in literary studies [50, 51, 11]. Rather than focusing on certain canonical works, distant reading aims at providing another perspective to complement closer analyses of some works [50]. Previous computational literary studies have focused on literature written in English, e.g. to study the literary canon of English and American literature in the 19th and 20th century [52, 53], and the themes in the 19th century novel [54]. Moreover, few projects have used LD resources to connect heterogeneous sources for an enriched view. Addressing these gaps, the BookSampo LD opens a new Finnish perspective to computational literary studies.

The ideas behind the Sampo model used in our work have been explored and developed before in different contexts. For example, the notion of collaborative content creation by data linking is a fundamental idea behind the Linked Open Data Cloud movement⁴⁴ and has been developed also in various other settings, e.g., in ResearchSpace⁴⁵. The idea of providing multiple analyses and visualizations to a set of filtered search results has been used in other portals, such as the ePistolarium⁴⁶ [55] for epistolary data, and using multiple perspectives have been studied as an approach in decision making [56]. Faceted search [57, 58], known also as view-based search [59, 60] and dynamic taxonomies [61], is a well-known paradigm for explorative search and browsing [62] in computer science and information retrieval, based on S. R. Ranganathan's original ideas of faceted classification in Library Science in the 1920's [63]. The two step study model used in our work has been used, e.g., in prosopographical research [64] (without the faceted search component). The novelty of the Sampo Model lies in combining several ideas and operationalizing them for developing LOD services and applications in Digital Humanities.

8. Conclusions and Future Work

The new UI offers the users a new and more intuitive way to explore the BOOKSAMPO KG compared to the old portal with its limited search functions. A key novelty in the new UI is the idea of integrating data-analytic tools seamlessly in the a library collection publishing system: the users can now also easily analyze the data they are presented without having to learn languages like Python to work with the result data from SPARQL queries. The underlying data offers possibilities for more specific queries and analyses for those who are able to query and process data. The new UI can serve as an intermediary step in DH research.

⁴²<https://ariadne-infrastructure.eu/>

⁴³<https://linked.art/>

⁴⁴<https://lod-cloud.net>

⁴⁵<https://www.researchspace.org>

⁴⁶<http://ckcc.huygens.knaw.nl>

The underlying BOOKSAMPO KG has a lot of potential to be used in literary DH research, but has not widely been used for it yet. With the new UI researchers could easily browse and explore the underlying data without needing to be familiar with technology like SPARQL or Python that would be needed otherwise. In addition to not needing to be familiar with the technology, the UI could be helpful in just narrowing down the area of interest to be researched, as finding interesting topics or questions could be easier with the UI than trying to scour through the BOOKSAMPO KG itself with its nearly 9 million triples.

The BOOKSAMPO data encompasses multiple aspects of the Finnish literature scene from novels to other formats of literature and contains rich metadata on these works. The data has a high potential for being used in Finnish literature research but has largely remained unused in spite of this. The data analysis carried out on the data strongly indicates that the data can indeed be used for DH research with success after doing some preprocessing to deal with the some of the problematic aspects of the data. The data analyses that were done as a part of this project was limited to just novels and a certain time period, leaving a lot of potential themes and subjects to be yet researched using the BOOKSAMPO data. In addition to this, the presented data analyses could benefit from combining it with some more close-reading approaches, e.g., by doing also text content analysis on some of the selected works from the works falling within the scope of the initial analysis.

The BOOKSAMPO 2.0 PORTAL could easily be extended in the future due to the nature of the Sampo-UI framework. The BOOKSAMPO KG includes data on several types of works, e.g., poems and short story collections, that could easily be added to the portal as perspectives after assessing the quality of that data and the annotations related to them. The components and visualizations present in the portal could be expanded as well based on the needs and wishes of users and their feedback.

Another avenue for further development is to apply to the model presented and software openly available to publishing and analyzing other literary datasets. According to our experiences [10] on developing over 20 Sampo portals [7], the declarative Sampo-UI framework can easily be attached to virtually any SPARQL endpoint and be adapted to the underlying data model and ontologies used.

Acknowledgements Thanks to Matti Sarmela, Kaisa Hypén, and Tuomas Aitonurmi for their collaborations as well as providing a newer version of the BOOKSAMPO KG to be used in the development of the new UI and the data analyses. Fruitful discussions and guidance on DH research of Mikko Tolonen and Eetu Mäkelä of the University of Helsinki, Department of Digital Humanities and HELDIG centre, are acknowledged. This project was funded by the Aalto University. The last author is thankful for the Cultural Foundation of Finland for an Eminentia Grant on reflecting research on the Sampo systems. Computing resources provided by the CSC – IT Center for Science were used in our work.

References

- [1] E. Mäkelä, K. Hypén, E. Hyvönen, Fiction Literature as Linked Open Data – the BookSampo Dataset, *Semantic Web – Interoperability, Usability, Applicability 4* (2013) 299–306. doi:10.3233/SW-120093.

- [2] E. Mäkelä, K. Hypén, E. Hyvönen, BookSampo—lessons learned in creating a semantic portal for fiction literature, in: *The Semantic Web – ISWC 2011*, Springer, 2011, pp. 173–188. doi:10.1007/978-3-642-25093-4_12].
- [3] E. Hyvönen, A. Ahola, E. Ikkala, Booksampo fiction literature knowledge graph revisited: Building a faceted search interface with seamlessly integrated data-analytic tools, in: *Theory and Practice of Digital Libraries (TDPL 2022)*, Accelerating Innovations Track, Padova, Italy, Springer, 2022. doi:10.1007/978-3-031-16802-4_54.
- [4] A. Ahola, E. Hyvönen, Visualizing literary linked data for public library users in the new user interface for booksampo – finnish fiction literature on the semantic web, in: *VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023*, CEUR Workshop Proceedings, Vol. 3508, 2023. URL: <https://ceur-ws.org/Vol-3508/paper1.pdf>.
- [5] E. Hyvönen, E. Mäkelä, T. Kauppinen, O. Alm, J. Kurki, T. Ruotsalo, K. Seppälä, J. Takala, K. Puputti, H. Kuittinen, K. Viljanen, J. Tuominen, T. Palonen, M. Frosterus, R. Sinkkilä, P. Paakkarinen, J. Laitio, K. Nyberg, CultureSampo – Finnish culture on the Semantic Web 2.0. Thematic perspectives for the end-user, in: *Museums and the Web 2009*, Archives & Museum Informatics, Toronto, 2009. URL: <https://www.archimuse.com/mw2009/papers/hyvonen/hyvonen.html>.
- [6] E. Mäkelä, T. Ruotsalo, Hyvönen, How to deal with massively heterogeneous cultural heritage data—lessons learned in CultureSampo, *Semantic Web – Interoperability, Usability, Applicability 3 (2012)* 85–109. doi:10.3233/SW-2012-0049.
- [7] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, *Semantic Web – Interoperability, Usability, Applicability 14 (2023)* 729–744. doi:10.3233/SW-190386.
- [8] E. Hyvönen, How to create a national cross-domain ontology and linked data infrastructure and use it on the Semantic Web, *Semantic Web – Interoperability, Usability, Applicability (2024)*. doi:10.3233/SW-2010-0014, in press.
- [9] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces, *Semantic Web – Interoperability, Usability, Applicability 13 (2022)* 69–84. doi:10.3233/SW-210428.
- [10] H. Rantala, A. Ahola, E. Ikkala, E. Hyvönen, How to create easily a data analytic semantic portal on top of a SPARQL endpoint: introducing the configurable Sampo-UI framework, in: *VOILA! 2023 Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2023*, CEUR Workshop Proceedings, Vol. 3508, 2023. URL: <https://ceur-ws.org/Vol-3508/paper3.pdf>.
- [11] F. Moretti, *Distant Reading*, Verso Books, 2013.
- [12] E. Gardiner, R. G. Musto, *The Digital Humanities: A Primer for Students and Scholars*, Cambridge University Press, New York, NY, USA, 2015. <https://doi.org/10.1017/CBO9781139003865>.
- [13] T. Koltay, Data literacy for researchers and data librarians, *Journal of Librarianship and Information Science* 49 (2015) 3–14. doi:10.1177/0961000615616450.
- [14] P. Hitzler, M. Krötzsch, S. Rudolph, *Foundations of Semantic Web technologies*, Springer, 2010.
- [15] T. Heath, C. Bizer, *Linked Data: Evolving the Web into a Global Data Space (1st edition)*,

- Morgan & Claypool, Palo Alto, California, 2011. URL: <http://linkeddatabook.com/editions/1.0/>.
- [16] E. Hyvönen, Publishing and using cultural heritage linked data on the Semantic Web, Morgan & Claypool, Palo Alto, California, 2012.
- [17] L. Rietveld, R. Hoekstra, The YASGUI family of SPARQL clients, *Semantic Web – Interoperability, Usability, Applicability* 8 (2017) 373–383. doi:10.3233/SW-150197.
- [18] P. Riva, M. Doerr, M. Zumer, FRBRoo: enabling a common view of information from memory institutions, in: *World Library and Information Congress: 74th IFLA General Conference and Council*, 2008.
- [19] E. Hyvönen, J. Tuominen, M. Alonen, E. Mäkelä, Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets, in: *ESWC 2014: The Semantic Web: ESWC 2014 Satellite Events*, Springer, 2014, pp. 226–230. doi:10.1007/978-3-319-11955-7_24.
- [20] A. Ahola, Developing a tool for information retrieval and research purposes utilizing BookSampo data, Master’s thesis, Aalto University, Department of Computer Science, 2023. URL: <https://urn.fi/URN:NBN:fi:aalto-202303262592>.
- [21] E. Ikkala, J. Tuominen, J. Raunamaa, T. Aalto, T. Ainiala, H. Uusitalo, E. Hyvönen, Namesampo: A linked open data infrastructure and workbench for toponomastic research, in: *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities, Geo-Humanities’18*, ACM, New York, NY, USA, 2018, pp. 2:1–2:9. doi:10.1145/3282933.3282936.
- [22] T. Burrows, N. B. Pinto, M. Cazals, A. Gaudin, H. Wijsman, Evaluating a semantic portal for the “Mapping Manuscript Migrations” project, *Digitalia* 15 (2020) 178–185.
- [23] E. Hyvönen, E. Ikkala, M. Koho, J. Tuominen, T. Burrows, L. Ransom, H. Wijsman, Mapping manuscript migrations on the semantic web: A semantic portal and linked open data service for premodern manuscript research, in: *Semantic Web. Proceedings of the The 20th International Semantic Web Conference (ISWC 2021)*, Springer, 2021, pp. 615–630. doi:10.1007/978-3-030-88361-4_36.
- [24] M. Hearst, A. Elliott, J. English, R. Sinha, K. Swearingen, K.-P. Lee, Finding the flow in web site search, *CACM* 45 (2002) 42–49.
- [25] J. English, M. Hearst, R. Sinha, K. Swearingen, K.-P. Lee, Flexible search and navigation using faceted metadata, Technical Report, University of Berkeley, School of Information Management and Systems, 2003.
- [26] T. Peura, Suomeksi yli rajojen. kvantitatiivinen tutkimus suomenkielisten romaanien monimuotoisuudesta 1970-2020, 2023. MSc Thesis.
- [27] P. Casanova, Literature as a world, *New Left Review* 31 (2005) 71–90. URL: <https://newleftreview.org/issues/ii31/articles/pascale-casanova-literature-as-a-world>.
- [28] G. Sapiro, Globalization and cultural diversity in the book market: The case of literary translations in the US and in France 38 (2010) 419–439. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0304422X1000032X>. doi:10.1016/j.poetic.2010.05.001.
- [29] M. Pollari, H.-L. Nissilä, K. Melkas, O. Löytty, R. Kauranen, H. Grönstrand, National, transnational and entangled literatures: Methodological considerations focusing on the case of finland (2015) 2–29. Cambridge Scholars Publishing Newcastle upon Tyne.
- [30] E. H. Simpson, Measurement of diversity, *Nature* 163 (1949) 688–688. URL: <https://www>.

nature.com/articles/163688a0. doi:10.1038/163688a0.

- [31] L. Huhtala, Dekkari nosteessa, Suomen nykykirjallisuus I (2013) 290–301.
- [32] O. Olaru, The internationalization of sjöwall and wahlöö. a quantitative study of scandinavian noir., in: DHN, 2019, pp. 333–348.
- [33] V. Ruohonen, Nordic noir on vahva brändi, entä suomi-noir?, AVAIN-Kirjallisuudentutkimuksen aikakauslehti (2018) 130–139.
- [34] K. T. Hansen, A. M. Waade, Locating Nordic Noir, Springer, 2017.
- [35] F. Moretti, Graphs, maps, trees: abstract models for a literary history, Verso, 2007. URL: <https://hdl.handle.net/2027/heb08911.0001.001>.
- [36] V. Sisättö, Tieteis- ja fantasiakirjallisuus, in: H. Riikonen, U. Kovala, P. Kujamäki, O. Paloposki, S. Höyhty, A.-M. Latikka (Eds.), Suomennoskirjallisuuden historia II, Suomalaisen Kirjallisuuden Seura, 2007.
- [37] S. Isomaa, T. Lahtinen, Kotimaisen nykydystopian monet muodot, Joutsen/Svanen (2017) 7–16.
- [38] J. Korpua, Spekulaatiivinen fiktio valtakielen markkinoilla, in: J. Ojajarvi, N. Työlähti (Eds.), Maamme romaani. Esseitä kirjallisuuden vuosikymmenistä., volume 121 of *Nykykulttuurin tutkimuskeskuksen julkaisuja*, Jyväskylän yliopisto, 2017, pp. 291–310.
- [39] J. English, M. Hearst, R. Sinha, K. Swearingen, K. Lee, Flexible search and navigation using faceted metadata, Technical Report, Technical report, University of Berkeley, School of Information Management . . . , 2002.
- [40] E. Sevänen, Suomennoskirjallisuuden määrällisestä kehityksestä, in: H. Riikonen, U. Kovala, P. Kujamäki, O. Paloposki, S. Höyhty, A.-M. Latikka (Eds.), Suomennoskirjallisuuden historia. 2, Suomalaisen Kirjallisuuden Seura, 2007, pp. 12–22.
- [41] B. Haslhofer, A. Isaac, R. Simon, Knowledge graphs in the libraries and digital humanities domain, arXiv preprint arXiv:1803.03198 (2018).
- [42] S. Van Hooland, R. Verborgh, Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata, Facet Publishing, 2014. doi:10.1080/00048623.2016.1162277.
- [43] M. Hallo, S. Luján-Mora, A. Maté, J. Trujillo, Current state of linked data in digital libraries, *Journal of Information Science* 42 (2016) 117–127. doi:10.1177/0165551515594729.
- [44] E. T. Mitchell, Library linked data: early activity and development, ALA TechSource Chicago, IL, 2016.
- [45] V. A. Carriero, A. Gangemi, M. L. Mancinelli, L. Marinucci, A. G. Nuzzolese, V. Presutti, C. Veninata, ArCo: The italian cultural heritage knowledge graph, in: *The Semantic Web – ISWC 2019*, Springer, 2019, pp. 36–52. doi:10.1007/978-3-030-30796-7_3.
- [46] Y. Lei, V. Lopez, E. Motta, V. Uren, An infrastructure for semantic web portals, *Journal of Web Engineering* 6 (2007) 283–308. URL: <https://journals.riverpublishers.com/index.php/JWE/article/view/4105>.
- [47] E. Hyvönen, K. Viljanen, J. Tuominen, K. Seppälä, Building a National Semantic Web Ontology and Ontology Service Infrastructure – The FinnONTO Approach, in: *Proceedings of the ESWC 2008*, Tenerife, Spain, Springer, 2008, pp. 95–109. doi:10.1007/978-3-540-68234-9_10.
- [48] M. Zeng, C. Sula, K. Gracy, E. Hyvönen, V. M. A. Lima, JASIST special issue on digital humanities (DH), *Journal of the Association for Information Science and Technology*

- (JASIST) (2021) 1–5. doi:10.1002/asi.24584.
- [49] H. R. Leo Lahti, Jani Marjanen, M. Tolonen, Bibliographic data science and the history of the book (c. 1500–1800), *Cataloging & Classification Quarterly* 57 (2019) 5–23. URL: <https://doi.org/10.1080/01639374.2018.1543747>. doi:10.1080/01639374.2018.1543747. arXiv:<https://doi.org/10.1080/01639374.2018.1543747>.
- [50] F. Moretti, Conjectures on world literature, *New left review* 2 (2000) 54–68.
- [51] T. Underwood, A genealogy of distant reading., *DHQ: Digital Humanities Quarterly* 11 (2017).
- [52] M. Algee-Hewitt, S. Allison, M. Gemma, R. Heuser, F. Moretti, H. Walser, Canon/archive: large-scale dynamics in the literary field, Technical Report, Stanford Literary Lab, 2016. URL: <https://litlab.stanford.edu/assets/pdf/LiteraryLabPamphlet11.pdf>.
- [53] R. Heuser, L. Le-Khac, A quantitative literary history of 2,958 nineteenth-century British novels: The semantic cohort method, Technical Report, Stanford Literary Lab, 2012.
- [54] M. L. Jockers, D. Mimno, Significant themes in 19th-century literature, *Poetics* 41 (2013) 750–769.
- [55] W. Ravenek, C. van den Heuvel, G. Gerritsen, The ePistolarium: Origins and techniques, in: A. van Hessen, J. Odiijk (Eds.), *CLARIN in the Low Countries*, Ubiquity Press, 2017, pp. 317–323. doi:10.5334/bbi.
- [56] H. A. Linstone, Multiple perspectives: Concept, applications, and user guidelines, *Systems practice* 2 (1989) 307–331. doi:10.1007/BF01059977.
- [57] M. Hearst, Design recommendations for hierarchical faceted search interfaces, in: *ACM SIGIR workshop on faceted search*, Seattle, WA, 2006, pp. 1–5.
- [58] D. Tunkelang, *Faceted search*, Morgan & Claypool Publishers, CA, USA, 2009.
- [59] A. S. Pollitt, The key role of classification and indexing in view-based searching, Technical Report, University of Huddersfield, UK, 1998. URL: <http://www.ifla.org/IV/ifla63/63polst.pdf>.
- [60] E. Hyvönen, S. Saarela, K. Viljanen, Application of ontology techniques to view-based semantic search and browsing, in: *The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004)*, Springer, 2004. doi:10.1007/978-3-540-25956-5_7.
- [61] G. M. Sacco, Dynamic taxonomies: guided interactive diagnostic assistance, in: N. Wickramasinghe (Ed.), *Encyclopedia of Healthcare Information Systems*, Idea Group, 2005.
- [62] G. Marchionini, Exploratory search: from finding to understanding, *Communications of the ACM* 49 (2006) 41–46. doi:10.1145/1121949.1121979.
- [63] A. C. Ferreira, B. C. M. dos Santos Maculan, M. M. L. Naves, Ranganathan and the faceted classification theory, *TransInformação*, Campinas 29 (2017) 279–295. doi:10.1590/2318-08892017000300006.
- [64] K. Verboven, M. Carlier, J. Dumolyn, A short manual to the art of prosopography, in: *Prosopography approaches and applications. A handbook*, Unit for Prosopographical Research (Linacre College), 2007, pp. 35–70. doi:1854/8212.