Estonian Scientific Research Trends

Team members: Ekke Gross, Leonhard Schecher

Project repository:

Business understanding

Identifying our business goals

Background

The goal of our project is to create a dashboard that would showcase recent research trends of Estonian scientific facilities. We will focus on research articles that have an abstract in ETIS or are publicly available.
The Estonian Research Information System (ETIS) stores information about Estonian researchers, institutions, projects, and publications. Currently, gaining an overview of how research topics change between departments and over time means doing manual searches and one-off reports. This makes the whole process tedious and time-consuming. The dashboard would be a solution that provides the viewer with the option to view those trends in a simple and non-bothersome way. Although ETIS has its own page for research statistics, it only covers the number of publications in a research field within ETIS during a specific time period. We aim to achieve more accurate results with the help of LLM, which would enable the classification of articles by keywords, based on context.

Business goals
Our project aims to summarize Estonian research activity over the last five years based on ETIS publication data, identify which research topics and keywords are growing or declining in importance, and present these trends in an interactive dashboard.

Business success criteria
The project will be considered successful if:

- The dashboard can answer typical management questions, such as “Which topics in computer science have grown the fastest since 2020?” or “How have department X’s publication trends changed over time?”
- All articles in the last five years that have an abstract in ETIS are included in our database.
- In a manual evaluation of a sample of articles, at least 80% of LLM-generated keywords are judged as relevant by the team.
- The project code and documentation are publicly available on the GitHub repository.

Assessing our situation

Inventory of resources
People: two data-mining students (project team) and the course instructor.
Data: ETIS public publication records.
Software and hardware: Python, Jupyter notebooks, pandas, GitHub, and access to the ChatGPT API through the university. We may also need to build a simple SQL database.

Requirements, assumptions, and constraints

- We use ETIS for broader metadata collection, but may need to deploy other data collection tactics.
- Time horizon: publications from the last five years.
- Only publicly available or institutionally authorised ETIS metadata will be used; no sensitive personal data is needed.

Risks and contingencies

- **Limited or difficult exports from ETIS.** Contingency: start with smaller subsets (selected departments) and, if needed, manually download structured exports.
- **ChatGPT API limits or high latency.** Contingency: batch requests, cache results locally, and prepare a fallback keyword extraction method (e.g. TF-IDF/RAKE) for a subset of publications.
- **Inconsistent department names or missing metadata.** Contingency: maintain a mapping table for department codes and allow some records to be grouped under “Unknown/Other”.

Terminology
Key terms for this project include:

- *Publication*: a research output recorded in ETIS (article, conference paper, monograph, etc.).
- *Classification:* we plan on using the Frascati classification to categorize the publications.
- *Keyword*: a topical descriptor assigned by the LLM based on the title and, if available, the abstract.
- *Trend*: a change in the frequency of a keyword or topic over the five-year period.

Costs and benefits
Costs consist mainly of student time (roughly 60 total person-hours) and modest compute/API usage. Benefits include a reusable ETIS-based research trends database, an interactive dashboard that could be used by researchers, and experience with LLM-based text enrichment. If successful, the approach could be extended to longer time periods.

Defining our data-mining goals

Data-mining goals

1. Build a structured database from ETIS containing all Estonian research articles from the last five years, including identifiers, publication year, department, publication type, language, and available abstracts.

2. Use the ChatGPT API to generate one to five concise keywords for each publication, in English, based on its title and abstract.

3. Aggregate publications by year, department, and keyword/topic to compute counts and relative shares.

4. Develop an interactive dashboard that allows filtering by year, department, field, and keyword to visualise trends and compare units.

Data-mining success criteria

- For a validation sample, LLM-generated keywords accurately capture the main topic of the article in at least 80% of cases; any disagreements will be analyzed and documented.

- The dashboard can be operated by a non-technical user after a short demo, and typical views load in under five seconds.

- All data preparation and modelling steps are documented according to CRISP-DM deliverables (data collection, description, exploration, and data quality reports).

Data understanding

Gathering data

Outline data requirements
To achieve our goals, we require publication-level metadata from ETIS covering:

- Publication identifier, title, year, and publication type.

- At least one organisational unit (department/institute) per publication.

- Existing ETIS classifications, such as research field codes (frascatti) or publication categories.

- Abstracts, when available, are used to provide the LLM with sufficient context for keyword extraction.

We require the metadata for all Estonian scientific research articles from the last five years that could be imported into a Python dataframe.

Verify data availability
ETIS is the national system that collects information about Estonian research institutions, researchers, projects, and research results, including publications. Public searches and institutional access enable the export of publication lists, and ETIS also exposes a Web API for public data. For this project, we assume that:

- We can filter exports by institution and publication year.

- We are permitted to use this metadata for non-commercial research and educational purposes.
- We may need to explore additional data-collection methods, as ETIS might not provide sufficient abstract data.

If full exports prove impossible, we will limit ourselves to selected departments or specific publication types and document this limitation.

Define selection criteria

- Institutions: all publishers whose publications' metadata are available in ETIS.

- Time period: the last five years.

- Publication types:
    1. Scholarly articles indexed by Web of Science Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Emerging Sources Citation Index and/or indexed by Scopus (excluding chapters in books)
    2. Peer-reviewed articles in other international research journals with an ISSN code and an international editorial board, which are circulated internationally and open to international contributions
    3. Scholarly articles in Estonian and other peer-reviewed research journals with a local editorial board; peer-reviewed scientific articles in journals important for Estonian culture or scholarly articles in Akadeemia, Looming, Vikerkaar
    4. Articles/chapters in books published by the publishers listed in Annex (including collections indexed by the Web of Science Book Citation Index, Web of Science Conference Proceedings Citation Index, Scopus)

5. Articles/chapters in books published by the publishers not listed in Annex
6. Full articles in encyclopedias

- Status: only verified/confirmed publications to avoid duplicates and drafts.

- Language: all languages are accepted, but LLM processing will primarily target records with English or Estonian titles/abstracts.

Describing data

After an initial ETIS export, we will create a data description report.

This report will summarise:

- Number of publications overall and per year.
- Number of fields per record and their types (numeric, categorical, free text).
- Coverage of key variables such as department, research field, and abstract.

This report will help us verify that we have enough data volume and the right attributes for meaningful analysis.

Exploring data

Data exploration will rely on simple summaries and visualisations:

- Frequency tables of publications by year, department, publication type, and language.
- Counts of missing values for key fields (department, year, abstract, field code).
- Checks for obvious outliers (e.g., duplicate titles with identical metadata).

We will also manually examine a small random sample of records to understand how departments and fields are recorded in ETIS and whether existing ETIS keywords or classifications could be used as additional features or evaluation signals in the future.

Verifying data quality

We will compile a data quality report that documents both minor and major issues, along with potential remedies. Anticipated problems include:

- Duplicated publications due to multiple ETIS entries. Remedy: deduplicate using combinations of title, year, DOI, and author list.
- Sparse or missing abstracts will limit the quality of LLM keywords. Remedy: fall back to using titles only, search for other sources for the data, or flag such records for separate evaluation.
- Remove any unnecessary data before presenting it to the LLM.

If we encounter severe issues (for example, entire years missing for some departments), we will document the impact and, if necessary, narrow the scope of the analysis rather than

over-interpret poor-quality data.

Planning our project

Task list and estimated effort

	Task	Ekke (h)	Leonhard (h)
1.	Refine requirements, set up repo, review ETIS/CRISP-DM documentation	4	2
2.	Data extraction	4	6
3.	Data description, exploration, and data quality reporting	9	6
4.	Design and implement an LLM-based keyword extraction pipeline, plus evaluation on a sample	5	8
5.	Design and test the dashboard.	6	6
6.	Preparing the presentation/dashboard demo	2	2

Total (approx.): Ekke 30 h, Leonhard 30 h.

Methods and tools

We plan to use ETIS exports as the primary data source, Python with Pandas, and may also utilize SQL (SQLite/PostgreSQL) for data handling, as well as Jupyter notebooks for experimentation. The ChatGPT API will be used to generate keywords from titles/abstracts, which will be saved. The visualization process has not been fully planned yet; most likely, we will build a web app using JavaScript libraries. The dashboard would most likely be a website, and the plots will be generated using plotnine/Seaborn, but we haven't fully decided on this matter either. All code and documentation will be maintained in our GitHub repository, following the CRISP-DM phases and deliverables as guidance.