I^3 Open Innovation Dataset Index

iiindex.org

Agnes Cameron, agnescam@mit.edu

December 4, 2021

Building an index for innovation data

- A place to curate lists of innovation datasets, tools and projects
- Different projects fit into different repositories: Zenodo, Dataverse, BigQuery all have different merits
- Likewise, multiple curated lists exist (NBER, LensLabs, Google Patents Public Datasets), for different needs

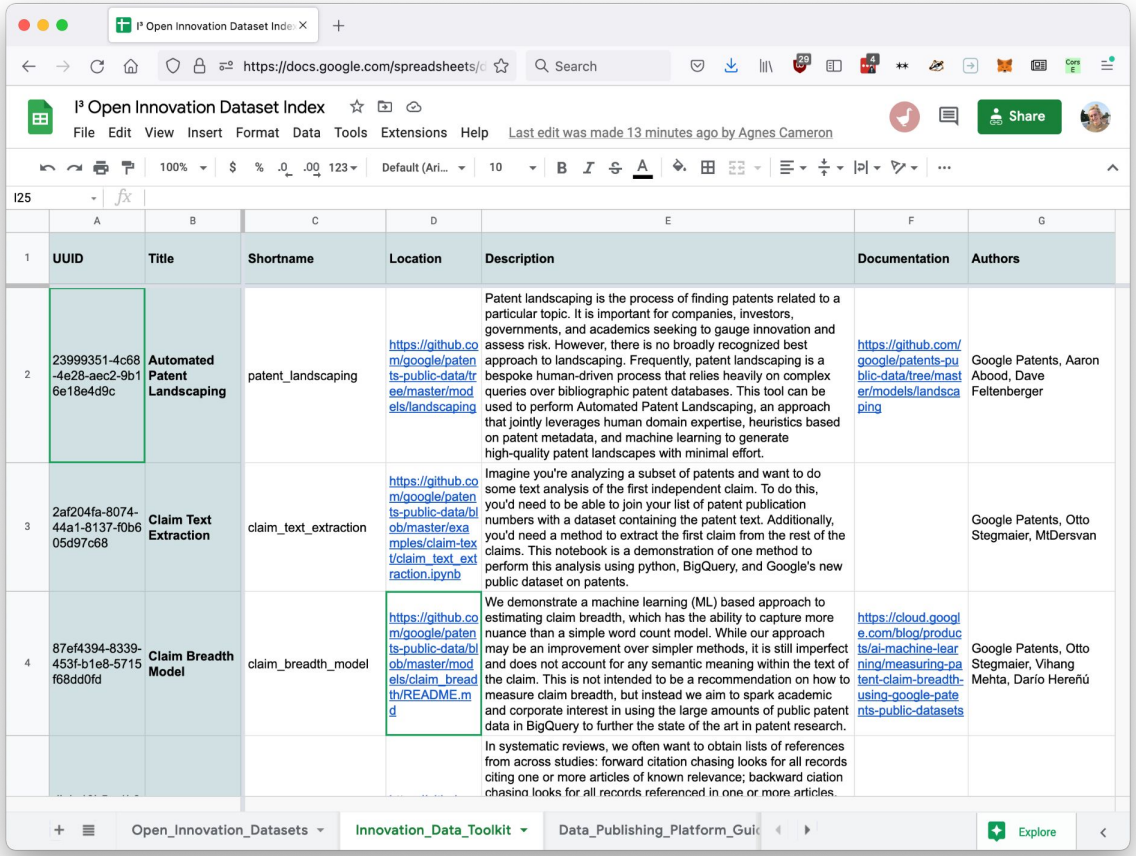
- The goal is not to replicate these, but create a lightweight, editable index to share + update annotated resources

    —> you can also add your own curated lists

Google BigQuery

First steps: Google Sheet

- Created Jan 2020
- Curated by I3 community
- Expanded to include tools and platforms
- Community editable
- Importantly: *still active*



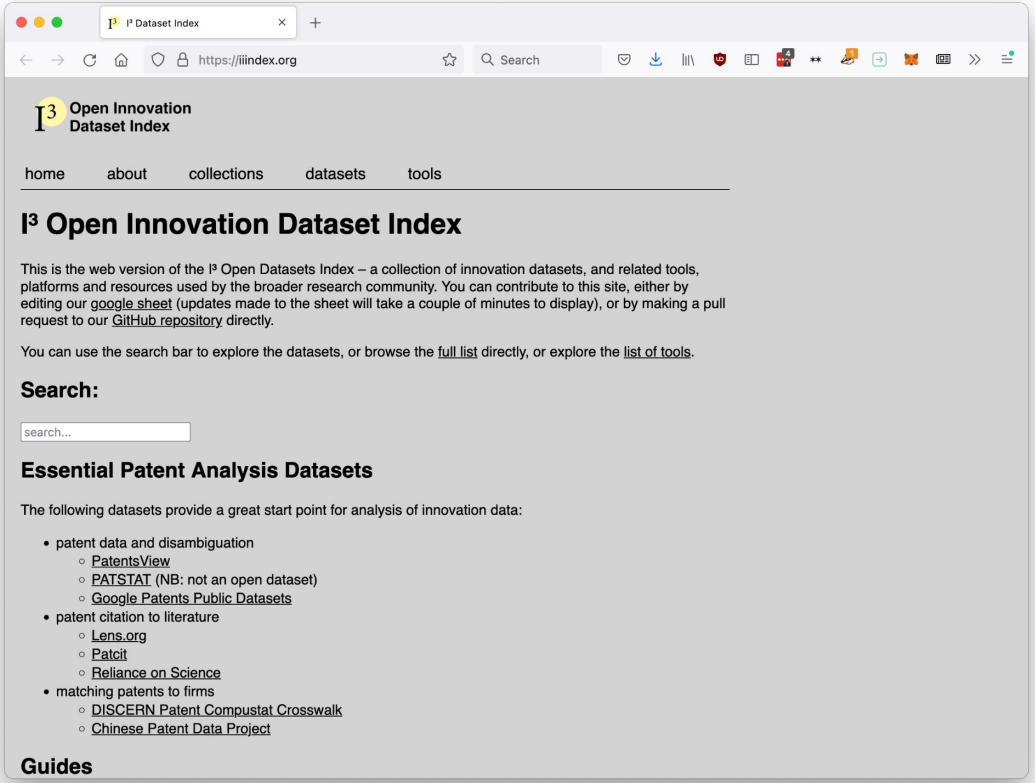
I³ Open Innovation Dataset Index

	UUID	Title	Shortname	Location	Description	Documentation	Authors
2	23999351-4c68-4e28-aec2-9b16e18e4d9c	**Automated Patent Landscaping**	patent_landscaping	https://github.com/google/patents-public-data/tree/master/models/landscaping	Patent landscaping is the process of finding patents related to a particular topic. It is important for companies, investors, governments, and academics seeking to gauge innovation and assess risk. However, there is no broadly recognized best approach to landscaping. Frequently, patent landscaping is a bespoke human-driven process that relies heavily on complex queries over bibliographic patent databases. This tool can be used to perform Automated Patent Landscaping, an approach that jointly leverages human domain expertise, heuristics based on patent metadata, and machine learning to generate high-quality patent landscapes with minimal effort.	https://github.com/google/patents-public-data/tree/master/models/landscaping	Google Patents, Aaron Abood, Dave Feltenberger
3	2af204fa-8074-44a1-8137-f0b605d97c68	**Claim Text Extraction**	claim_text_extraction	https://github.com/google/patents-public-data/blob/master/examples/claim-text/claim_text_extraction.ipynb	Imagine you're analyzing a subset of patents and want to do some text analysis of the first independent claim. To do this, you'd need to be able to join your list of patent publication numbers with a dataset containing the patent text. Additionally, you'd need a method to extract the first claim from the rest of the claims. This notebook is a demonstration of one method to perform this analysis using python, BigQuery, and Google's new public dataset on patents.		Google Patents, Otto Stegmaier, MtDersvan
4	87ef4394-8339-453f-b1e8-5715f68dd0fd	**Claim Breadth Model**	claim_breadth_model	https://github.com/google/patents-public-data/blob/master/models/claim_breadth/README.md	We demonstrate a machine learning (ML) based approach to estimating claim breadth, which has the ability to capture more nuance than a simple word count model. While our approach may be an improvement over simpler methods, it is still imperfect and does not account for any semantic meaning within the text of the claim. This is not intended to be a recommendation on how to measure claim breadth, but instead we aim to spark academic and corporate interest in using the large amounts of public patent data in BigQuery to further the state of the art in patent research.	https://cloud.google.com/blog/products/ai-machine-learning/measuring-patent-claim-breadth-using-google-patents-public-datasets	Google Patents, Otto Stegmaier, Vihang Mehta, Darío Hereñú
					In systematic reviews, we often want to obtain lists of references from across studies: forward citation chasing looks for all records citing one or more articles of known relevance; backward ciation chasing looks for all records referenced in one or more articles.		

I3 Open Innovation Index Website

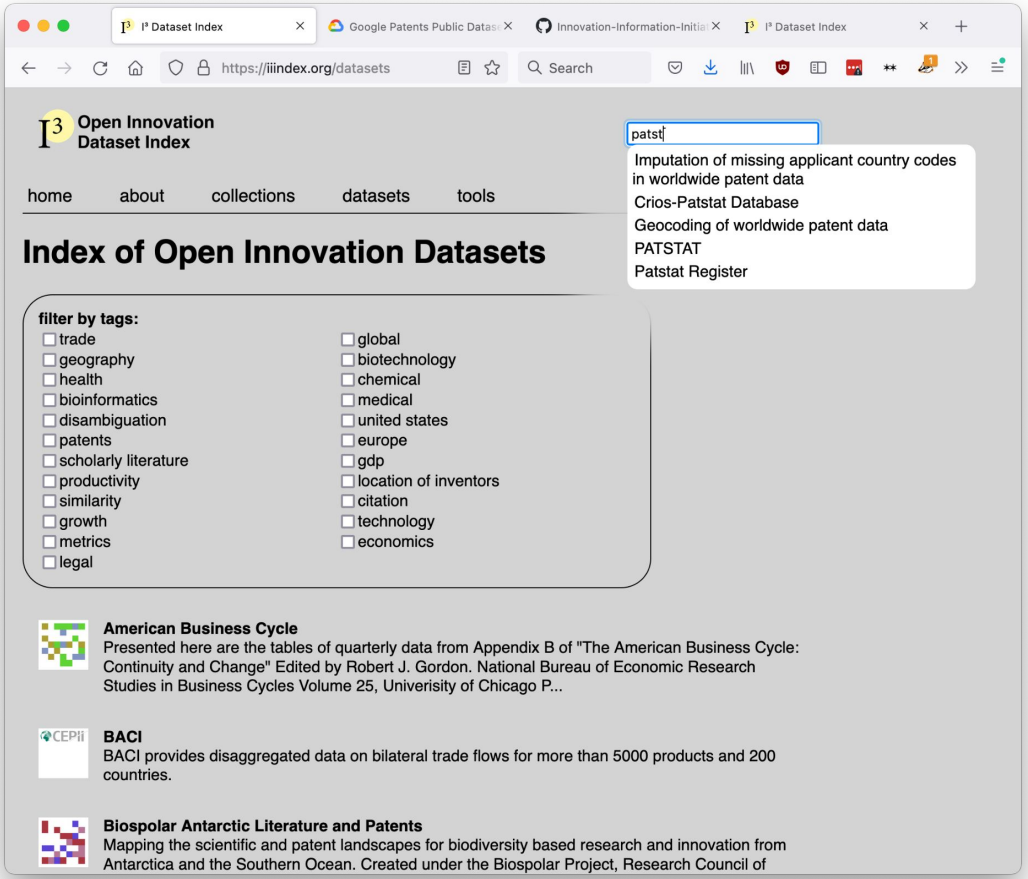
- *Still* editable through google sheet
- *Also* editable through github
- Open edits, full version history

- Searchable and linkable
- Allows for community-curated dataset indexes and guides
- Makes space for notes, annotations, links and examples



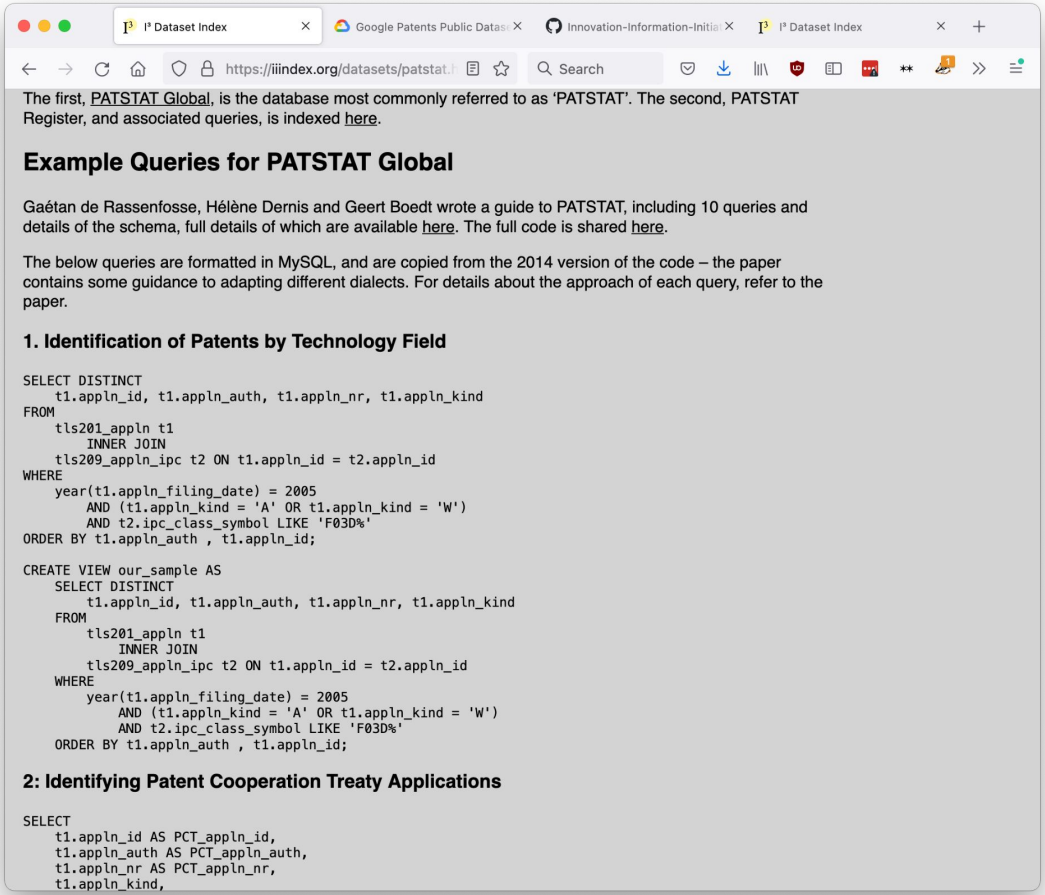
Searchable index

- Lists of tools and datasets
- Filter by common tags
- In the future: advanced search queries
  - By author
  - By timeline
  - By related datasets



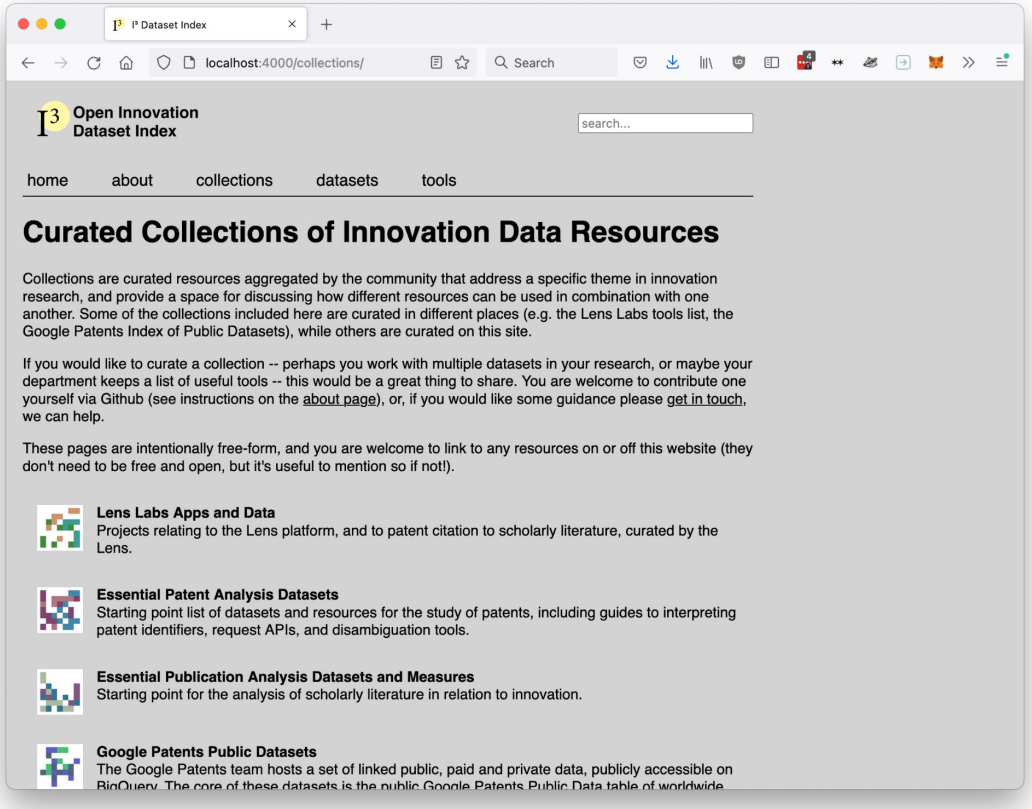
Annotatable files

- Each page in the index is an editable markdown file
- Headers include metadata from the google sheet
- Markdown body provides open-ended space to add notes and queries



Curated collections

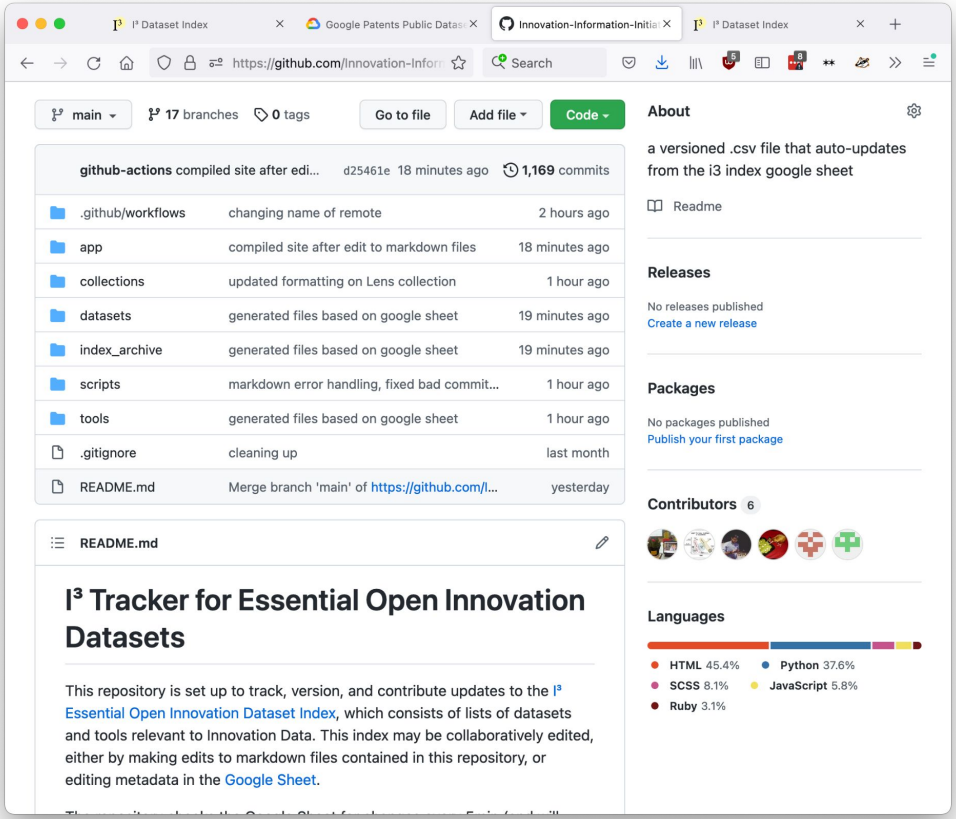
- An open way to list thematic data
- Some collections track external repositories (e.g. Lens Labs, NBER)
- **This is a great place to contribute!**



How to contribute

- Did you publish a dataset?
  → add it to the google sheet
- Do you use code to process a dataset?
  → add a sample query on Github
- Have you made a list of datasets for a project, or to give to your students?
  → create a collection

- **Guidelines for contributing** are on the [about page] and on the [github]



Future extensions

- Automated updates of metadata and versions (cron job that fetches metadata)
- Use APIs to get greater range of citation metadata, rather than just MediaWiki citations
- Expand use of ‘superceded by’ tag

Technical details

- Updates to the site are automated via Github actions
- Pull requests trigger scripts that update the information
- A versioned copy of each table in the google sheet is published in .csv form
- Citation metadata for datasets & tools is harvested automatically when submissions are made