



Digital Science Technical Report

This Poster is Reproducible

Building Reproducible Data Science Workflows with
Dimensions, Gigantum, Figshare, and Overleaf

Simon Porter and Jared Watts

SEPTEMBER 2019



About Digital Science

Digital Science is a technology company working to make research more efficient. We invest in, nurture and support innovative businesses and technologies that make all parts of the research process more open and effective. Our portfolio includes the admired brands Altmetric, Anywhere Access, CC Technology, Dimensions, Figshare, Gigantum, GRID, IFI Claims, Overleaf, ReadCube, Ripeta, Symplectic and Writefull. We believe that together, we can help researchers make a difference. Visit www.digital-science.com

About Dimensions

Dimensions is an innovative research knowledge system that re-imagines discovery and access to research. Developed by Digital Science in collaboration with over 100 leading research organizations around the world, Dimensions links grants, publications, citations, alternative metrics, clinical trials and patents. It enables users to find and access the most relevant information faster, analyze the academic and broader outcomes of research, and gather insights to inform future strategy. Data and expertise that span the research life cycle were contributed by the teams at Digital Science portfolio companies ReadCube, Altmetric, Figshare, Symplectic, Digital Science Consultancy and ÜberResearch, who came together to realize their unique strengths and share their passion for building tools that benefit the research community. Find out more at www.dimensions.ai

About Overleaf

With over 4.3M registered users, **Overleaf** is an academic authorship tool that allows seamless collaboration and effortless manuscript submission, all underpinned by cloud-technology. By providing an intuitive online collaborative writing and publishing platform, Overleaf is making the process of writing, editing and publishing scientific documents quicker and easier. Researchers and academics can now write, collaborate, and publish with a single click, directly from the Overleaf web app. Publishers and institutions are partnering with Overleaf to provide customized writing templates, simple reference tool linking, and one-click publishing submission links. Supported by Digital Science, Overleaf aims to make science and research faster, more open and more transparent by bringing the whole scientific writing process into one place in the cloud – from idea, to writing, to review, to publication. Visit www.overleaf.com

About Figshare

Figshare is a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner. Figshare's aim is to become the place where all academics make their research openly available. It provides a secure cloud based storage space for research outputs and encourages its users to manage their research in a more organized manner, so that it can be easily made open to comply with funder mandates. Openly available research outputs will mean that academia can truly reproduce and build on top of the research of others. Visit www.figshare.com

About Gigantum

Gigantum develops software that increases transparency & reproducibility in science and data science. We want people to directly access, inspect, understand, and build on each other's work without all of the technical and practical hurdles they face today. Through automation of best practices and thoughtful user interfaces we strive to level the playing field for all users. From novice to expert, individual with a laptop to a professional researcher with a large cloud compute budget, Gigantum can improve your daily workflow and enhance the transparency and reproducibility of your work. Visit www.gigantum.com

About the Authors

Simon Porter, Simon Porter came to Digital Science from the University of Melbourne, where he worked for past 15 years in roles spanning the Library, Research Administration, and Information Technology. Beginning from a core strength in the understanding of how information on research is collected, Simon has forged a career transforming university practices in how data about research is used, both from administrative and eResearch perspectives. In addition to making key contributions to research information visualization and discovery within the university, Simon is well known for his advocacy of Research Profiling Systems and their capability to create new opportunities for researchers. Over the past three years, Simon has established and run the annual Australasian conference on research profiling. In 2012, Simon was the program chair of the third annual VIVO conference.

 orcid.org/0000-0002-6151-8423

Jared Watts, Before starting at Digital Science, Jared Watts worked at the University of Auckland for the last ten years. Jared is a software engineer by trade, with his development experience spanning over twenty years. At the University of Auckland his area of focus included library management systems, library discovery systems, research information management systems, institutional repositories and data preservation.

 orcid.org/0000-0002-3315-1572

This report has been published by Digital Science which is part of Holtzbrinck, a global media company dedicated to science and education.

DOI:10.6084/m9.figshare.9741890

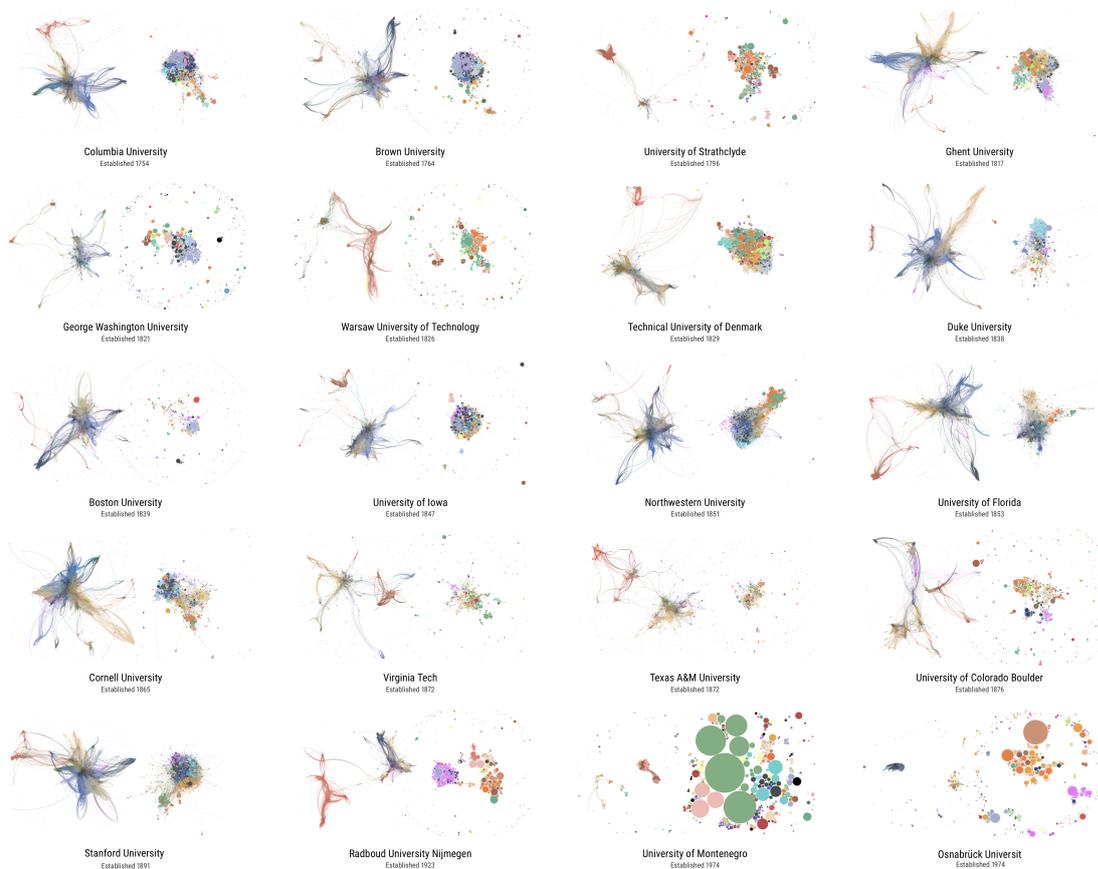
Digital Science, 6 Briset Street, London EC1M 5NR info@digital-science.com

Copyright: © 2019 Digital Science. This work is licensed under the Creative Commons Attribution 4.0 International License CC-BY.

What Does a University Look Like?

Simon Porter orcid.org/0000-0002-6151-8423, Jared Watts orcid.org/0000-0002-3315-1572, Digital Science, September 2, 2019

External and internal co-authorship network diagrams provide different perspectives on the collaborative shapes of research institutions. The external (left) and internal (right) collaboration patterns of the universities of VIVO attendees are presented here. Researchers are coloured by the field of research that they most commonly publish in, and sized by total number of publications that they have published (relative to the network). To create the networks, publications published between Jan 2015 and July 2019 were analysed.



Field of Research Colour Codes

Pure Mathematics	Soil Sciences	Environmental Engineering	Other Medical and Health Sciences
Applied Mathematics	Biochemistry and Cell Biology	Geomatic Engineering	Curriculum and Pedagogy
Numerical and Computational Mathematics	Ecology	Manufacturing Engineering	Specialised Studies in Education
Statistics	Evolutionary Biology	Maritime Engineering	Economic Theory
Mathematical Physics	Genetics	Materials Engineering	Applied Economics
Astronomical and Space Sciences	Microbiology	Resources Engineering and Extractive Metallurgy	Econometrics
Atomic, Molecular, Nuclear, Particle & Plasma Phys	Physiology	Interdisciplinary Engineering	Banking
Condensed Matter Physics	Plant Biology	Medical Biotechnology	Policy and Administration
Optical Physics	Zoology	Communications Technologies	Political Science
Quantum Physics	Other Biological Sciences	Nanotechnology	Sociology
Other Physical Sciences	Animal Production	Medical Biochemistry and Metabolomics	Psychology
Analytical Chemistry	Crop and Pasture Production	Cardiorespiratory Medicine and Haematology	Law
Inorganic Chemistry	Fisheries Sciences	Clinical Sciences	Film, Television and Digital Media
Macromolecular and Materials Chemistry	Forestry Sciences	Dentistry	Cultural Studies
Organic Chemistry	Horticultural Production	Human Movement and Sports Science	Linguistics
Physical Chemistry (incl. Structural)	Veterinary Sciences	Immunology	Archaeology
Theoretical and Computational Chemistry	Artificial Intelligence and Image Processing	Medical Microbiology	Curatorial and Related Studies
Other Chemical Sciences	Computation Theory and Mathematics	Neurosciences	Historical Studies
Atmospheric Sciences	Computer Software	Nutrition and Dietetics	Applied Ethics
Geochemistry	Data Format	Oncology and Carcinogenesis	History and Philosophy of Specific Fields
Ecology	Information Systems	Ophthalmology and Optometry	Philosophy
Geophysics	Biomedical Engineering	Pediatrics and Reproductive Medicine	Religion and Religious Studies
Oceanography	Chemical Engineering	Pharmacology and Pharmaceutical Sciences	
Physical Geography and Environmental Geoscience	Civil Engineering	Medical Physiology	
Environmental Science and Management	Electrical and Electronic Engineering	Public Health and Health Services	

Figure I: 'What does a university look like?' poster

Contents

1	Introduction	5
2	The Subject Matter: What does a University look like?	6
2.1	Using Dimensions to Extract Research Network Information	6
3	Scaling up the Production of Network Diagrams	7
4	Working in Gigantum	8
5	Using Figshare with Gigantum	9
6	Making a Poster in Overleaf	10
7	Making an interactive version of the Poster	11
8	Integrating Gigantum, Figshare, and Overleaf from the Beginning:	12

I Introduction

Expectations around reproducible research are clear, particularly in the area of computational research [11]. A research paper is more than an account of the research that was undertaken; it is a narrative that surrounds an orchestration of research assets from the raw data and code, to the processed data and visualizations that result. A paper should invite a reader to trace the results back. How was this figure produced? What was the code that produced that particular result? The reader's transition from narrative to exploring data or code should be as easy as turning the page.

Seen from the researcher's perspective, the ideal computational paper arises organically from the research - the data that is created **is** the data that ends up in the paper. The code as it is written **is** the code that can be accessed in the paper. As analysis bubbles up from research into images for publication, those images keep their providence back to the data, and back to the code that produced them.

How close are we to this ideal today? Within the Digital Science family, methods for openly publishing data are ably supported by Figshare. Gigantum provides researchers with productive environments to not only develop their code, but also share their projects along with the providence of the steps that were run, and the environment necessary to execute it. Overleaf allows researchers to easily publish their research collaboratively using LaTeX. As part of a poster presentation for the 2019 VIVO conference we took a broad research question that could be answered with Dimensions data, and undertook the research using workflows that knit these tools together. In doing so, this project demonstrates an approach to undertaking reproducible computational science that operates on multiple levels. Specifically, it addresses:

- What is it to develop reproducible of code right from the beginning of a project using Gigantum?
- How can data assets be structured and organized throughout the life of a project inside Figshare (and not just at the end of a project)?
- How can Overleaf be integrated with Gigantum so that the act of creating an image or data table is as close as possible to the act of publishing the same object in a paper?
- What is a good approach to tying code, data, and papers together using identifiers?

By the end of the paper we hope we will have demonstrated that not only is our poster reproducible, but that the methods we have adopted are useful to others as well.



Figure 2: Ghent University (FOR)



Figure 3: Ghent University (Gender)

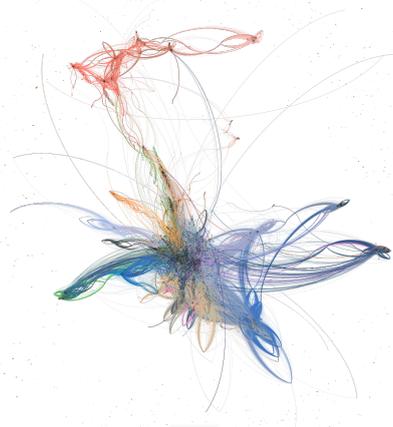


Figure 4: Columbia University (External FOR)

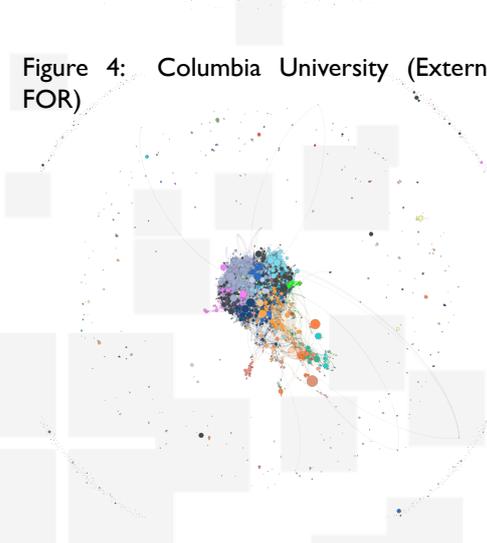


Figure 5: Columbia University (Internal FOR)

2 The Subject Matter: What does a University look like?

Universities come in many different shapes and sizes, but how different are their collaboration footprints? Via its API, the Dimensions data model[5] can be used to readily reveal information on the demographics of researchers, and the different ways that they collaborate. Late last year, Digital Science released an interactive dashboard looking at the gender distribution of researchers by institution and by field of research[10]. Further examples that demonstrate how relatively easy it is to extract this information from the API have also been recently published.[8]

Collaboration networks derived from co-authorship information on publications are an effective way to visualize researcher demographic patterns as they place individual demographic indicators in the context of research communities. Different types of demographic information can then be overlaid on the network to determine the shape and size of the nodes. By looking at the entire publication history of a researcher, it is possible to identify the Field of Research that they most commonly publish in (fig.2).

Although a researcher will publish in multiple fields over a career, identifying a researcher's most common field of research is a good way to identify the perspective that a researcher brings to a collaboration. The size of a researcher's node can be calculated based on the researcher's total number of career publications. Using the grid[9] data that is associated with each author affiliation, we can establish a researcher's likely home institution and country. We can also establish a researcher's likely gender (fig.3) based on their first name and the use of a gender guessing tool[7]. (currently, the utility of this approach is limited to western names)

Depending on whether external researchers were included in the graph, two different perspectives emerge:

- A network graph that indicates external researcher highlights an institutions integration into global research networks
- A network graph that focuses on internal researcher highlights how ideas are exchanged internally across different areas of research

2.1 Using Dimensions to Extract Research Network Information

To build a collaboration network for an entire university over a given time period, we start with a Dimensions query like this:

```
search publications
  where year in [2015:2019]
    and research_orgs.id = "a grid id"
    and count(researchers) < 400
    and type = "article"
return publications[id + author_affiliations]
```

A grid id is used to limit the results to a particular institution. We chose to limit the number of researchers to 400 as publications with greater than 400 authors tend to dominate the shape of any resulting network diagram, and significantly increase the time it takes to visualize the resulting network graph. (The ability to filter results by the number of authors is one of my favourite features of the Dimensions API)

Having retrieved the publications we are interested in, we created a network of researchers based on their co-authorship relationships as expressed in the author affiliations section of the publication record.

For the size of each node we chose to represent the number of career publications for each researcher. For the colour of each node, we chose to represent the four digit Field of Research (FOR) code to which their research was most commonly associated. We also created two alternate network graphs with gender

4 Working in Gigantum

From the beginning, the code for this project was developed in Gigantum. Gigantum offers several advantages over working directly in a Jupyter notebook environment such as Anaconda. Specifically, Gigantum:

- sets up your project from the very beginning in way that makes it easy to share your code, and all of its dependencies
- tracks of of your changes
- allows you to share your code with collaborators
- makes your code portable, so you can shift your code from machine to machine depending on the resources (CPU, storage, stability, etc) that you need

When you set up a Gigantum project, you explicitly specify the dependencies that your code required at both the python level as well as the operating level (such as java and vncserver for running gephi.) Gigantum manages each project in a separate docker container so each project is running in its own environment. At the end of your project, publishing and storing your code is as simple as publishing your project.

With the exception of adding your own API passwords etc, running your project will be as straight forward for you as anyone else that chooses to run your project.

For this project, running Gigantum on a server instance was particularly useful as many of the steps took an overnight run to complete, and were not suitable for running on a laptop.

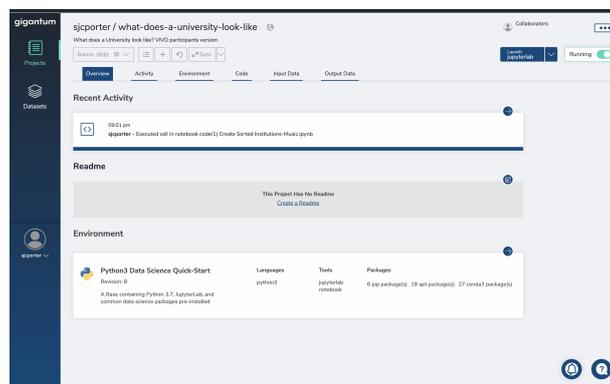


Figure 7: Gigantum project interface

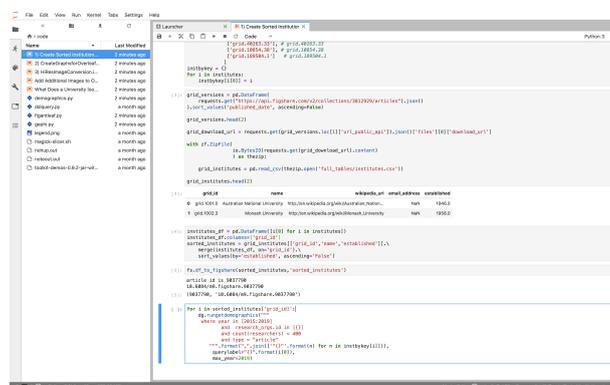


Figure 8: Running Jupyter Notebooks within Gigantum

5 Using Figshare with Gigantum

A revolution in education and technology transfer follows from the marriage of word processing and software command scripts. In this marriage an author attaches to every figure caption a push button or a name tag usable to recalculate the figure from all its data, parameters, and programs... ...preparing such electronic documents is little effort beyond our customary report writing; mainly, we need to file everything in a systematic way - Jon F. Claerbout and Martin Karrenbach. 1992[2]

A challenge of working on a data project inside a Docker container (via Gigantum) is that Docker can slow down and use up a lot of resources if there are many large files in your project (especially on a Mac). In addition, large files, (or indeed many small files,) can significantly slow down synchronising your project changes to git. To get around this, we chose to shift data assets to Figshare as soon as they were created.

Structuring the data in Figshare, was just as important as storing it. In order to manage the data that would be eventually made public, we chose to create one Figshare article per university that we analysed. Each 'university' article consisted of:

- a file containing the anonymized researcher definitions.
- a file containing the the relationships between researchers and publications.
- a version of the relationships file that only contained the internal relationships (these were stored as csv, and pickle files for easy retrieval into a pandas data frame)

The article also stored:

- the resulting graphml file that was produced from both the internal and external versions of the graphs.
- the gephi files that were created as a result of the custom gephi jar file.
- graph images at different resolutions and formats for different media presentations. For the poster presentation we created png images at 1024. pixels. For online exploration, we created images at a significantly larger size (40960 pixels). These larger images were converted from png files to dzi format using magick-slicer [12] so that they could be rendered in openseadragon [3].

By collecting the files together in this way, a single figshare article becomes an interface to an object that not only allows access the data behind the graph, but also its representations for different media. The structure of the figshare file also allows for code rerun at different levels. For example, if you don't have access to the Dimensions API, the code can be run from the original data files instead of creating them afresh with the API.

In addition to the Figshare articles that were produced for each University, companion private Figshare articles were created that contain the identified raw data from Dimensions. (lightly anonymized data was used for this project as the purpose of the outputs was not to assess individual researcher contributions.

To manage the 40+ figshare articles that form part of this exercise, a Figshare project that corresponded to the Gigantum project was created. The Figshare project also offered another level of flexibility in that the data assets could be shared with collaborators, allowing different parts of the project (code and data) to be shared independently.

Shifting data at the point of creation from Gigantum to Figshare had another benefit. Gigantum keeps a record of all code that has been executed, so by reserving a doi at the point of data creation, and including it in the execution log, data can be tied to the exact code sequence that created it.

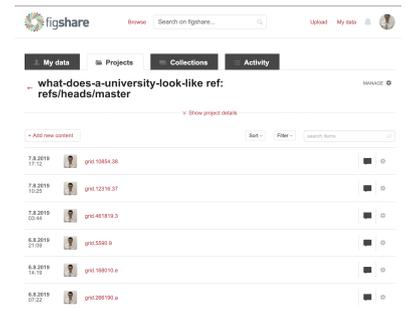


Figure 9: Articles produced in Gigantum - managed in figshare

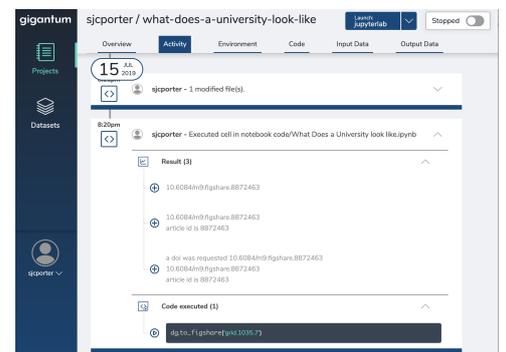


Figure 10: Gigantum activity log, with link to figshare

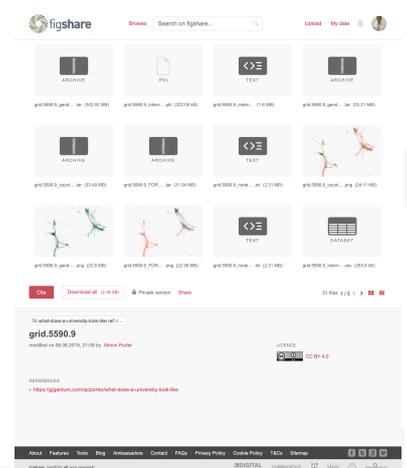


Figure 11: Figshare article, with link to the Gigantum project that created it in the references

6 Making a Poster in Overleaf

For data driven documents, collaborating on presentation layer activities (such as writing a paper or poster based on emerging results) is challenging. Typically, each result object such as a table or image is copied, rather than linked into the document. With the data embedded, every time the data is updated all images and tables based on that data need to be re copied into the document. In most cases, this means a whole lot of manual work. The consequence of this process is that writing up the paper has to come as close to the end of the project as possible in order to avoid a lot of wasted manual effort.

Writing a data driven document is far more like building a website, where the text and images are linked together from different files, any of which could be updated separately to each other. And like a modern website, many different people can work on multiple aspects of the presentation all at once.

Writing a document in LaTeX shares many similarities to building a website. Documents are created in code that are then rendered into different formats (typically PDF.) Like html, there are tools that make writing LaTeX documents much easier (such as Overleaf.) Overleaf is particularly suited to producing data driven documents as an Overleaf project can be synchronised to Github, where other tools can be used to push and update data and images into the project. [1]

To create the VIVO poster, at the same time that files were pushed to Figshare, we also pushed a print version of the graph to a Github repository that was integrated with an Overleaf project. As each image was created with its own DOI, a bibtex file containing the references to the figshare dois, and a supplementary.tex file with the images embedded for easy copying and pasting into the main tex document was also programatically produced. The original workflow envisaged that each image would link to its own doi citation in Figshare, however for reasons of real estate, we chose to reference the links to the figures in the online version instead.

Within the Overleaf project, each graph image was given a sequential image name (fig1, fig2 etc,) as well as a corresponding fig_X_caption.tex file that contained its description. In this way the Overleaf template created could remain independent of the specific University details that were pushed to it via the github integration. To make a poster with a different set of Universities, it should be sufficient to copy the tex file containing the poster layout file into another project.

To keep track of the files that were already pushed to Overleaf, we created a overleaf.csv file that was stored as a separate article inside the Figshare project.

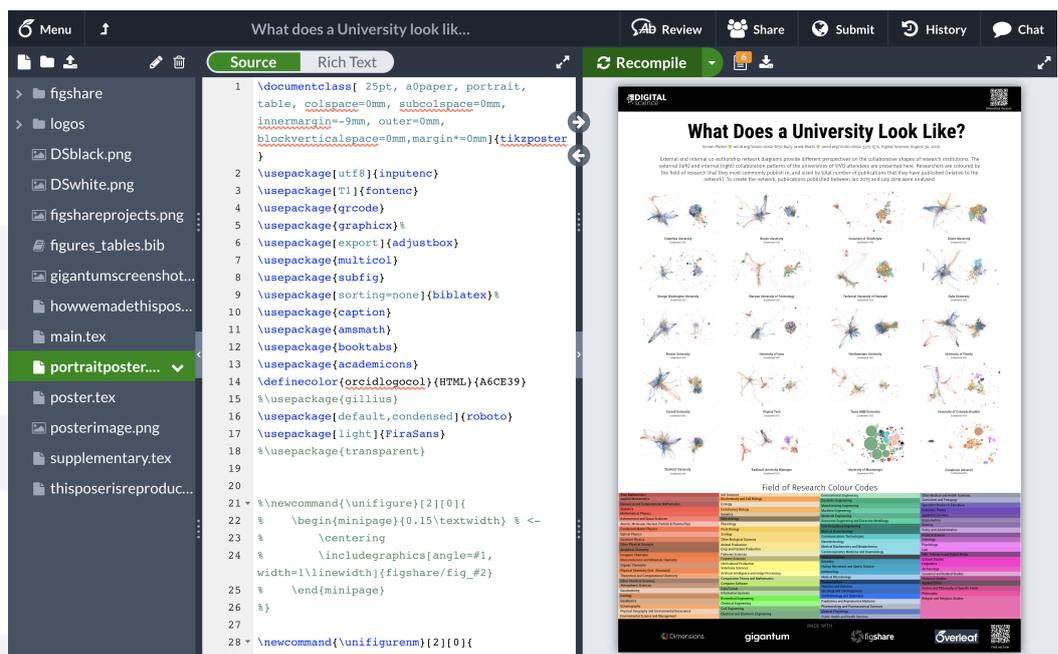


Figure 12: Using Overleaf to make the poster

7 Making an interactive version of the Poster

A limitation of the poster format (no matter how big) is that detailed network diagrams can take up a lot of space on the page. To compliment the poster, we produced an online version of the University network graphs that allow users to interact with the networks, and zoom in to a much higher resolution. To enable graphs to be rendered quickly, we processed the large graph outputs with magick-Slicer [12] to convert them into dzi format. The dzi images were then rendered with openseadragon [3]. Node interactivity was enabled via an additional file created as part of the gephi graph generation process detailing the coordinates of each node. These coordinates were then mapped onto the image. In the same way as the Overleaf poster was designed to work with multiple institutions as inputs, the Online site works directly with the figshare project, and renders each of the University files in the project as inputs.

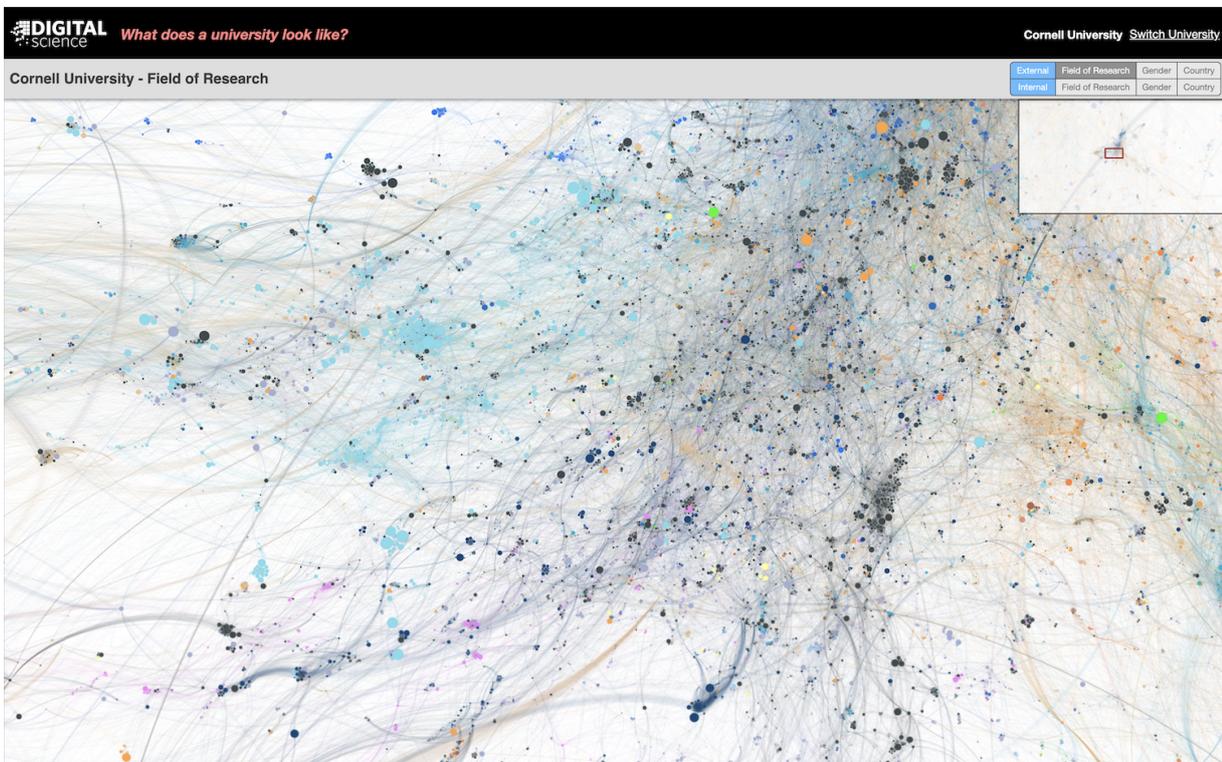


Figure 13: Online Version

8 Integrating Gigantum, Figshare, and Overleaf from the Beginning:

The pattern of integrating Gigantum, Figshare, and Overleaf together from the beginning is one that should benefit all data driven projects. To that end, we have refactored the code that synchronizes Gigantum with Figshare and Overleaf into a separate library that can be imported into any Jupyter project that runs in a Gigantum environment. The python module is called (unimaginatively) `figantleaf`.

The `figantleaf` module embodies a number of principles:

1. Every table or image in a paper (or poster) should have its own DOI
2. Every article in figshare should act as an interface to multiple representations of the data
3. Preparing an image for citation should be as easy as indicating that you wish to include that image in your code
4. Files belonging to a Gigantum project should be organised as a figshare project

Although not visible in the output of the Poster, `Figantleaf` has a number of features.

1. Using the function `df_to_figshare` will publish a pandas dataframe to Figshare in two formats, a 'pickle' file containing the binary representation of the dataframe, and a csv version for human readable discovery, and a doi will be reserved.
2. Using the function `fig_to_figshare` will publish a matplotlib figure to Figshare, in two resolutions, one for print and one for the web. It will also include the dataframe on which the figure was based by calling `df_to_figshare`. At the same time `fig_to_figshare` can be used to push the print image to Overleaf
3. Using the function `graph_to_figshare` with two dataframes supplying the nodes and edges data will result in an figshare article that stores the graphml file, as well as the two original dataframes in pickle and csv format. In addition to `graph_to_figshare`, a separate `gephi` module was developed to assist with creating and pushing graph images to figshare and Overleaf.

Data from the Figshare project can be called back into the Jupyter notebook by calling `figshare_to_df`, with either the name of the article, and optionally specific filename. The same approach is used for `figshare_to_graph`. In this way data assets can be used across Jupyter notebooks as easily as if they were stored in the local file system.

Setting up `figantleaf` to run at the start of a Gigantum project involves a few minutes of setup. A Figshare token is required to integrate Gigantum with Figshare, and the name of the Github repository that has been created by Overleaf has to be supplied. Additionally ssh keys for Github need to be setup. The `figantleaf` module will prompt the user through the steps required.

As the additional functionality provided in `figantleaf` indicates, the pattern of connecting gigantum to figshare and overleaf has become more than just an exercise in using Digital Science products from end to end. Particularly when using easy to access data sources such as the Dimensions API, we feel a workflow that uses Gigantum for data Analysis, Figshare for data management and publication, and Overleaf for presentation is a workflow that can be used effectively today, and one that we hope to make even easier to do in the future.

References

- [1] Arin Basu. *Data science workflow with Jupyter and Overleaf v2, Part I: setup*. 2018. URL: <https://arinbasu.svbt1e.com/data-science-workflow-with-jupyter-and-overleaf-v2-part-i-setup> (visited on 08/13/2019).
- [2] Jon F. Claerbout and Martin Karrenbach. “Electronic documents give reproducible research a new meaning”. In: *SEG Technical Program Expanded Abstracts 1992*. 1992, pp. 601–604. DOI: 10.1190/1.1822162. URL: <https://app.dimensions.ai/details/publication/pub.1098913318>.
- [3] OpenSeadragon contributors CodePlex Foundation. *MagickSlicer*. 2009. URL: <https://openseadragon.github.io> (visited on 08/28/2019).
- [4] gephi.org. *Gephi Toolkit*. 2010. URL: <https://gephi.org/toolkit/> (visited on 08/13/2019).
- [5] Daniel W. Hook, Simon J. Porter, and Christian Herzog. “Dimensions: Building Context for Search and Evaluation”. In: *Frontiers in Research Metrics and Analytics* 3 (2018), p. 23. DOI: 10.3389/frma.2018.00023. URL: <https://www.frontiersin.org/articles/10.3389/frma.2018.00023/pdf>.
- [6] Mathieu Jacomy et al. “ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software”. In: *PLoS ONE* 9.6 (2014), e98679. DOI: 10.1371/journal.pone.0098679. URL: <https://app.dimensions.ai/details/publication/pub.1048491501>.
- [7] Israel Saeta Pérez. *Pypi gender-guesser 0.4.0*. 2016. URL: <https://pypi.org/project/gender-guesser/> (visited on 08/13/2019).
- [8] Simon Porter. *Dimensions API + Jupyter notebooks examples*. Nov. 2018. DOI: 10.6084/m9.figshare.c.4262975.v4. URL: https://figshare.com/collections/Dimensions_API_Jupyter_notebooks_examples/4262975/4.
- [9] Digital Science. *GRID Global Research Identifier Database*. 2015. URL: <http://grid.ac> (visited on 08/13/2019).
- [10] Digital Science et al. *Gender Representation in UK Research Institutions*. Jan. 2019. DOI: 10.6084/m9.figshare.7583402.v1. URL: https://digitalscience.figshare.com/articles/Gender_Representation_in_UK_Research_Institutions/7583402/1.
- [11] Victoria Stodden et al. “Enhancing reproducibility for computational methods”. In: *Science* 354.6317 (2016), pp. 1240–1241. DOI: 10.1126/science.aah6168. URL: <https://app.dimensions.ai/details/publication/pub.1001771541>.
- [12] VoidVolker. *MagickSlicer*. 2015. URL: <https://github.com/VoidVolker/MagickSlicer> (visited on 08/28/2019).

Part of **DIGITAL**science

