Contrasting platforms and infrastructures as configurations for data sharing

Jean-Christophe Plantin, LSE

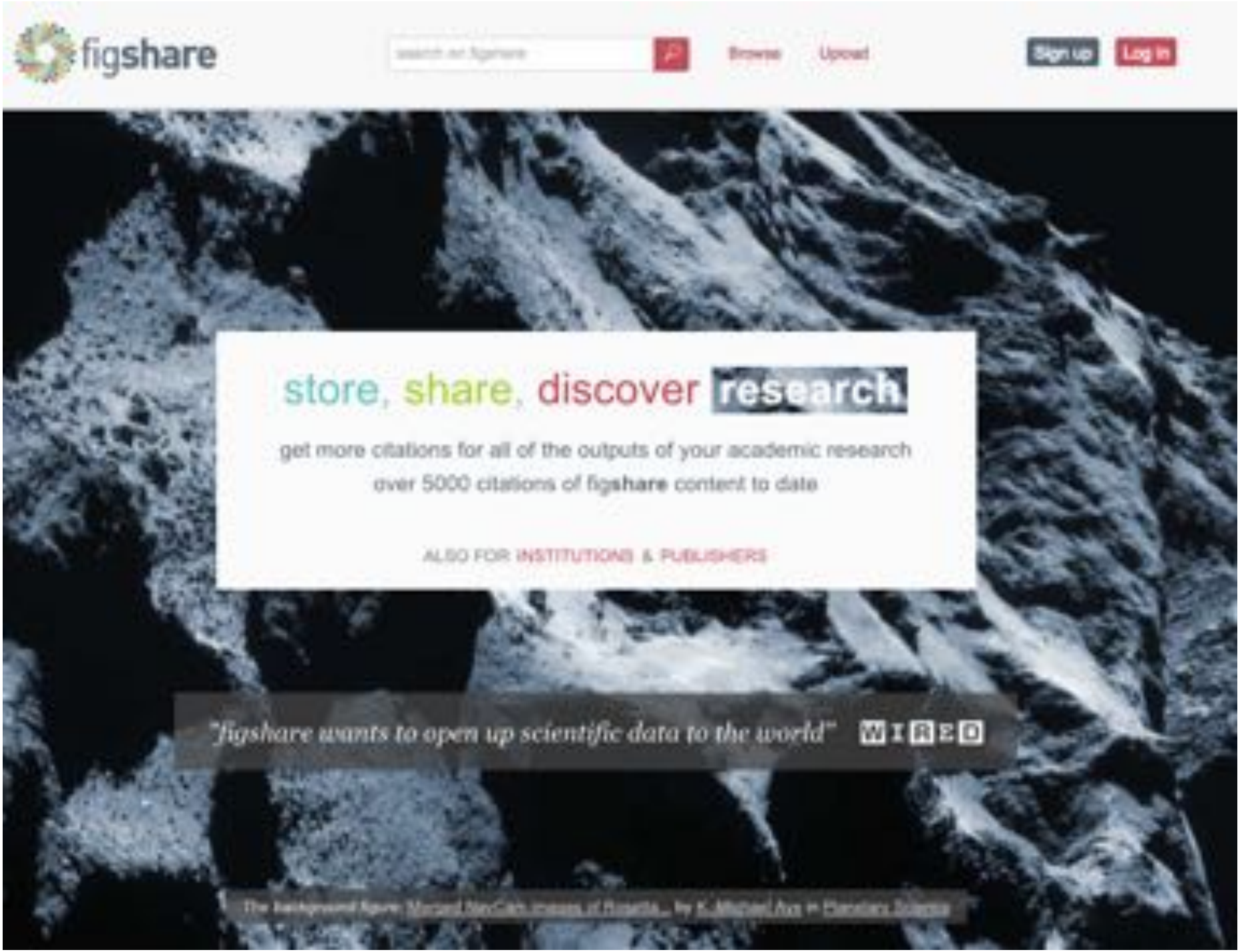
DARIAH's Humanities at Scale Winter School in Prague: 24th-28th October 2016
Session 09 Infrastructures & platforms

Big Data, data as scholarly output

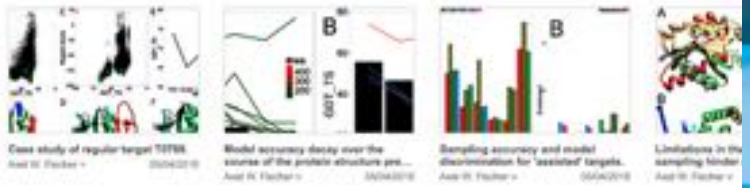
- Big data across disciplines
- New data sharing requirements
- Materials considered as scholarly outputs

The decentralisation of scholarly infrastructure

- World wide web!
- e-print movement and the institutional repository effort of the early 2000's
- E.g. ArXiv.org


figshare
Browse
Upload
Sign up
Log in
store, share, discover research
get more citations for all of the outputs of your academic research
over 5000 citations of figshare content to date
ALSO FOR INSTITUTIONS & PUBLISHERS
"figshare wants to open up scientific data to the world"
WIRED


figshare
Browse
Upload
Sign up
Login
PLOS
Discover research from Public Library of Science -
NEW
POPULAR
CATEGORIES
SEARCH



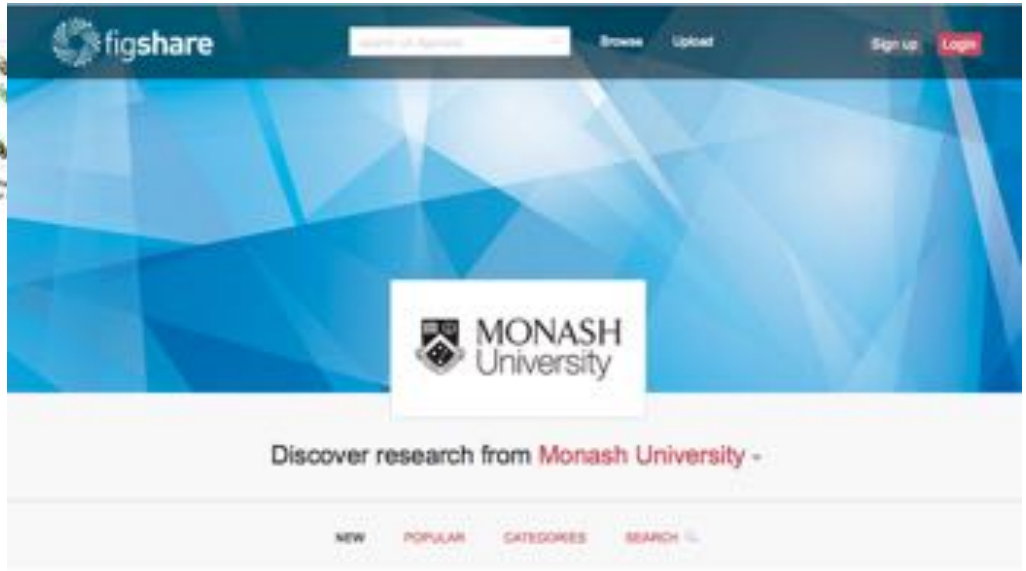

figshare
Browse
Upload
Sign up
Login
MONASH
University
Discover research from Monash University -
NEW
POPULAR
CATEGORIES
SEARCH

Table 1. Table summarizing infrastructure and platform properties.

	Infrastructure	Platform
Architecture	Heterogeneous systems and networks connected via sociotechnical gateways	Programmable, stable core system; modular, variable complementary components
Relation between components	Interoperability through standards	Programmability within affordances, APIs
Market structure	Administratively regulated in public interest; sometimes private or public monopoly	Private, competitive, sometimes regulated via antitrust and intellectual property
Focal interest	Public value; essential services	Private profit, user benefits
Standardization	Negotiated or de facto	Unilaterally imposed by platforms
Temporality	Long-term sustainability, reliability	Frequent updating for competitive environment
Scale	Large to very large; ubiquitous, widely accessible	Small to very large; may grow to become ubiquitous
Funding	Government, subscription, lifeline services for indigent customers, pay-per-use (e.g. tickets)	Platform purchase (device), subscription (online), pay-per-use (e.g. TV shows), advertising
Agency of users	"Opt out," for example, going off the grid	"Opt in," for example, choosing one platform instead of another; creating mashups

Plantin, Jean-Christophe, Carl Lagoze, Paul N. Edwards, and Christian Sandvig. "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook." *New Media & Society*, August 4, 2016

Table 1. Table summarizing infrastructure and platform properties.

	Infrastructure	Platform
Architecture	Heterogeneous systems and networks connected via	Programmable, stable core system; modular, variable
Relation between components	Interoperability through standards	Programmability within affordances, APIs
Market structure	Administratively regulated in public interest; sometimes private or public monopoly	Private, competitive, sometimes regulated via antitrust and intellectual property
Focal interest	Public value; essential services	Private profit, user benefits
Standardization	Negotiated or de facto	Unilaterally imposed by platforms
Temporality	Long-term sustainability, reliability	Frequent updating for competitive environment
Scale	Large to very large; ubiquitous, widely accessible	Small to very large; may grow to become ubiquitous
Funding	Government, subscription, lifeline services for indigent customers, pay-per-use (e.g. tickets)	Platform purchase (device), subscription (online), pay-per-use (e.g. TV shows), advertising
Agency of users	"Opt out," for example, going off the grid	"Opt in," for example, choosing one platform instead of another; creating mashups

Plantin, Jean-Christophe, Carl Lagoze, Paul N. Edwards, and Christian Sandvig. "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook." *New Media & Society*, August 4, 2016

Inter-university Consortium for Political and Social Research

ICPSR is an international resource for scientific cooperation. Its purpose is to maximize the availability of social science data and minimize the inconvenience and cost of teaching and studying social and historical phenomena.

Since 1962 ICPSR has collected over 5000 computer-readable data files on social phenomena, both contemporary and historical, and made them conveniently available to Consortium members at a reasonable cost. From its original twelve members interested mainly in political science research, the membership has grown to more than 240 institutions sharing the world's largest collection of computer-readable data in all fields of social science.

This vast archive provides data on social phenomena in more than 130 countries — surveys of mass and elite attitudes, census records, election returns, international relations, and legislative records. The data on individual and social experience extend across the full range of the social sciences — history, sociology, social psychology, economics, geography, political science, public health, public policy, education, and international studies.

All of these basic research resources are available to ICPSR member institutions in formats compatible with local computing facilities.

ICPSR regularly consults with members on the application of the most advanced computer software technology to social science research. The Consortium also conducts annual training programs in the theory and technology of social research.

Social science research and instruction have been significantly refined by the thousands of scholars who have made use of ICPSR's training and data resources.

ICPSR

ICPSR—a resource for data, training and computing assistance in the social sciences

Care of data through manual processing

Action	1. Deposit the dataset	2. Dispatch	3. Repair		4. Contact with the PI (optional)	5. Prepare		6. Verify	7. Publish
Description	The PI or acquisition department deposits a study for processing	The manager reviews and dispatches the study to a processor	The processor first reviews the data, identifies problems, and draws a processing plan	The processor then "fixes" the problems: "wild codes," missing values, questions, labels, etc.	The processor, after contact with the manager, contacts the PI	The processor writes the metadata for the study	The processor formats the datasets and the documents according to ICPSR templates	The processor sends all the files to a manager and another processor for "Quality check"	Once reviewed, the manager approves the publication of the study on the ICPSR website
Staff	Principal Investigator (PI)	Manager	Processor		PI/Processor/ Manager	Processor		Processor/Manager	Processor
Tool	Deposit form	ICPSR internal workspace	Scripts, Unix, notepad, SPSS, "eyeball"		Email, spreadsheet	PDF, Hermes, ICPSR internal workspace		Unix, PDF, notepad, SPSS	ICPSR internal workspace

Table 1. Reconstitution of the "pipeline" for data processing

(Plantin, forthcoming)

Figshare: Automatic data provision

- No processing, self-deposit
- Centrality of the API
  - To connect web-based actors with the scholarly world
  - To connect with institutions and publishers

Code as a Research Object

Mozilla Science Lab

Get credit for your code!

Archive your GitHub code repository to figshare and receive a citable DOI.

Start Now

GitHub

Edit, share and improve your code in a collaborative environment.

Mozilla Science Lab

Tools to get your research on the web.

figshare

Persistent, citable, long-term archiving for your research outputs.

Bridging the discourses of web development and science

For us the Shangri-La is a world where a group of researchers generate new information based on new hypotheses, make it available on Figshare and then others pull together the combined world's knowledge to look for new patterns and discoveries that the original authors had not thought to look for. These data scientists will need tools, which we have already made available on our open API. I'd love to see more of these tools helping more people make more discoveries, more often – Mark Hahnel, 2014

Consequences on scholarship

- Infrastructures:
  - Manual processing for specific type of data, but what about big heterogeneous data?
  - Path dependence and reverse salient (Hughes 1983)
- Platforms:
  - No means to guarantee figshare will remain open
  - Splintering infratructures (Graham, Marvin, 2001)

Contrasting platforms and infrastructures as configurations for data sharing

Jean-Christophe Plantin, LSE

DARIAH's Humanities at Scale Winter School in Prague: 24th-28th October 2016
Session 09 Infrastructures & platforms