

EMCODIST: Introducing the ‘Digitally Curious’ to Email Archives for Organizational Research and History



Update for
Virtual Email Archiving Workshop
June 13-15, 2023

Prof. David A. Kirsch
University of Maryland, College Park
Robert H. Smith School of Business
College of Information Studies





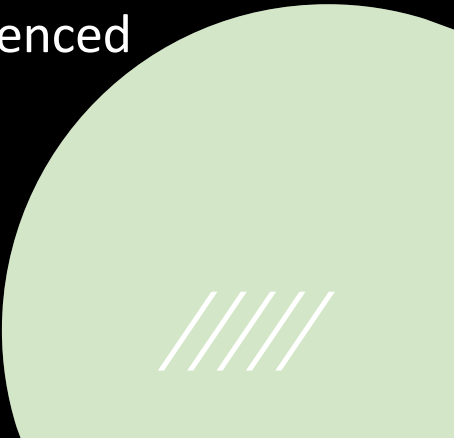
A user perspective

- Born-digital sources are vital for future historical research
- Many born-digital collections remain inaccessible*
- Ethical implications of using digital sources are unclear
- Many researchers trained in the use of physical, pre-digital sources





Beyond Preservation

- Born-digital sources are shaking up traditional archival processes (Prom et al., 2019)
 - Managing privacy at scale remains a key barrier to born-digital access (Milligan, 2019)
 - Answering the problem of access requires collaboration ‘between both sides of the reading room’ (Jaillant, 2019)
 - Users value optionality, expect some curation, but are still inexperienced (Wellcome Trust, 2017)
- 

How will users *actually* engage with born digital material once access issues have been navigated?



Contextualizing Email Archives

Explored the gap between current efforts to preserve emails, and the means by which researchers might actually read and engage with them.



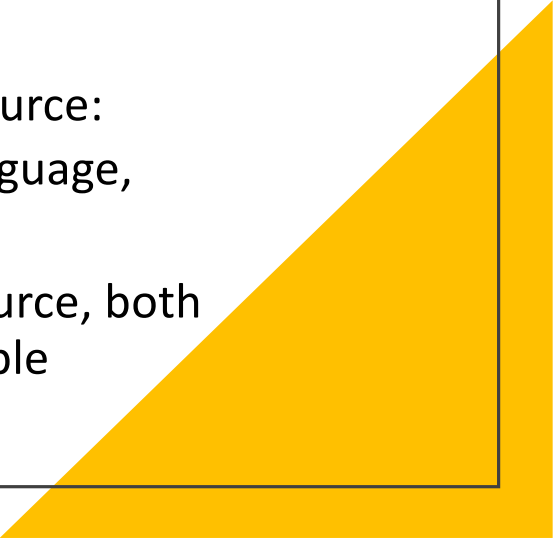
THE	
NATIONAL	
ARCHIVES	

- Used the emails of a failed US Dot-Com company
- Preserved and made available for research via LDC
- Collaboration with TNA digital archives specialists

Project tested new ways to make an email archive available to search and study while maintaining the basic relational and network properties of the format



Email specific issues

- Networked nature of organizational email make it difficult to search
 - Email is a hybrid artifact: email IS and email ARE
 - Not just information as content, but also as context
 - Non-historians often engage with just one aspect of resource:
 - Frequency and networks, timing and sequencing, language, content
 - For organizational email to become a useful historical source, both individual and network aspects of email must be accessible
- 

We also
assume a
need to...



Accommodate increasingly diverse
research questions



Allow users to work iteratively through a
collection



Work with the tacit (sometimes messy)
nature of historical research



Provide for different levels of experience



Offer relatively complete access to a
whole organisational corpus

Working Hypothesis

With a relatively complete e-mail archive, a scholar can ask and answer most important historical questions about an organization.**

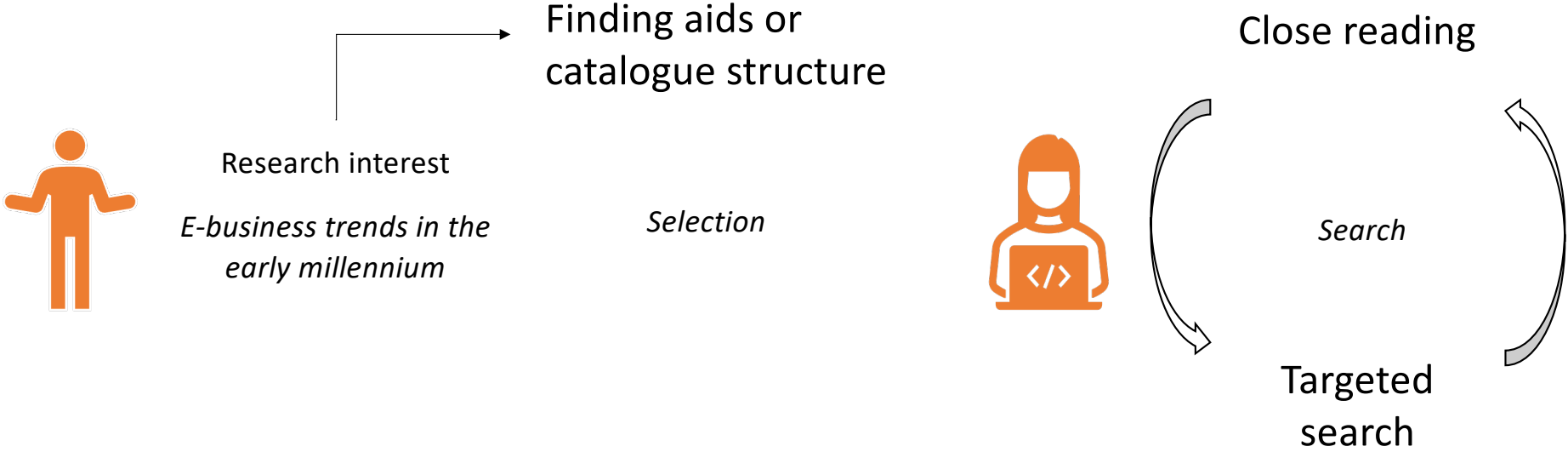


Users of born-digital

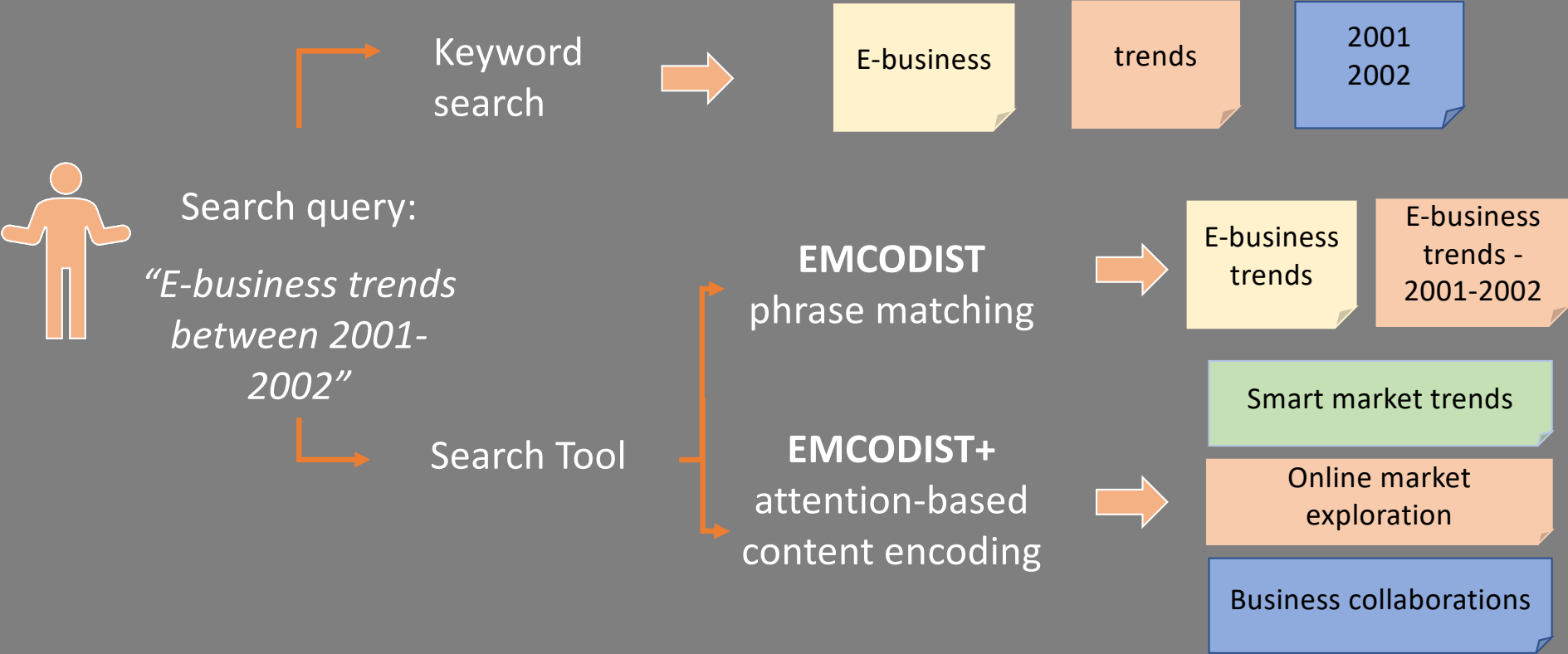
Talboom and Underdown (2019) provide a typology of users:

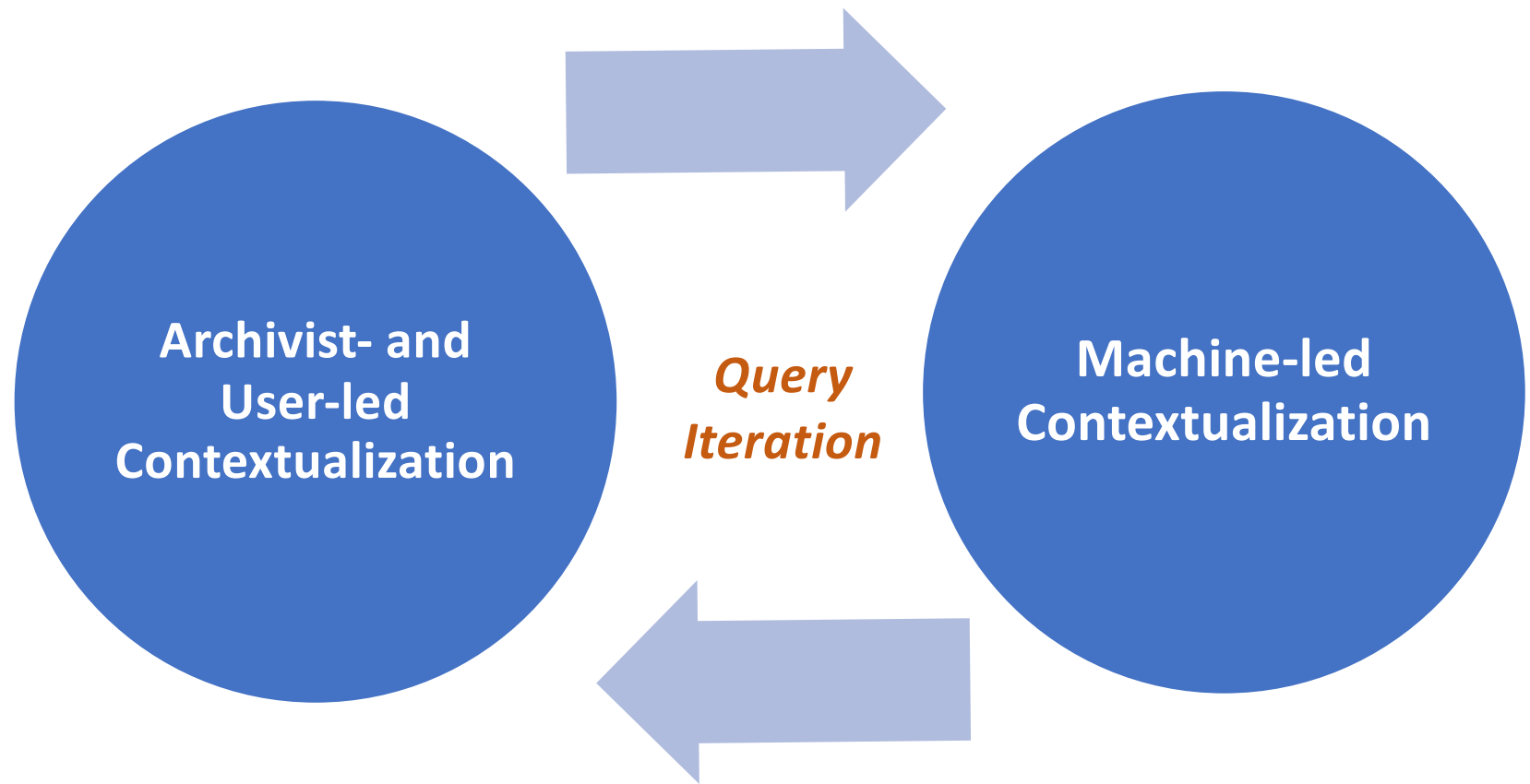
- “readers” want to access a digital source like a traditional paper source
- “digitally curious” want to search large databases to identify items of importance for more in-depth study
- “data users” want to perform computational analysis over entire collections.

Archivist- and User-led Contextualization



Prototype: Machine-led Contextualization







EMCODIST

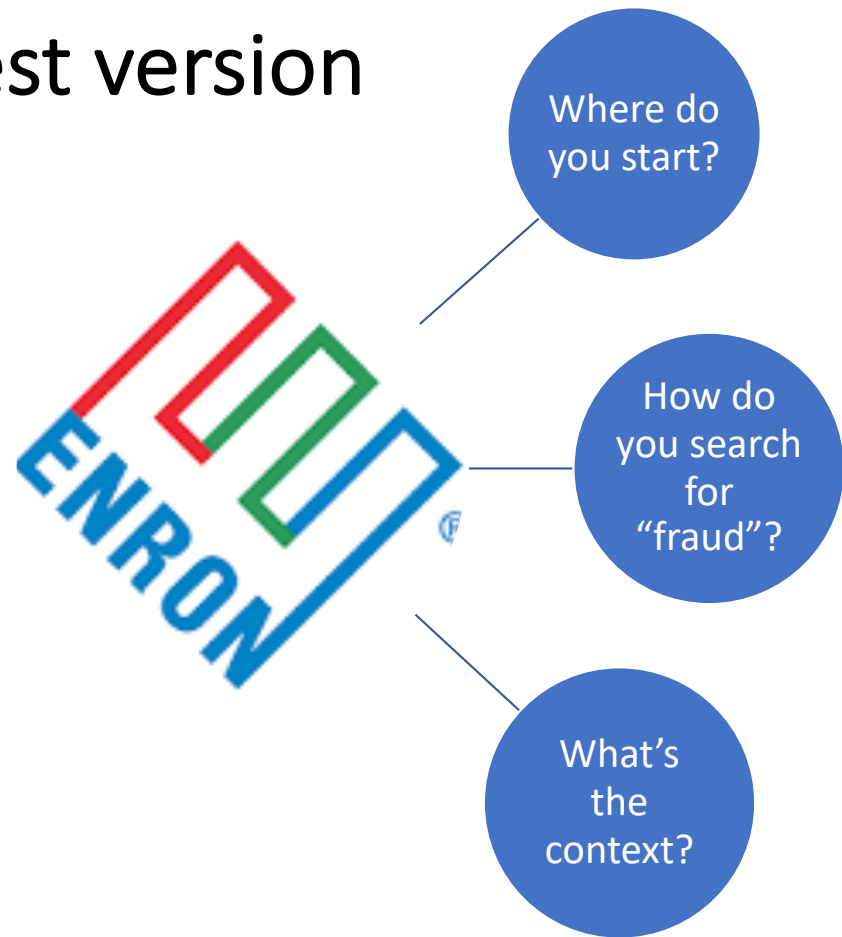
- Finding out how researchers actually use our tool
- Review user behaviour when using the tool (logging of activity)
- Request user evaluation through a short survey
- How do users navigate the empty search box?
 - 'Googlefication' of search
 - How to phrase a query when you do not know what is in the resource

Enron: the EMCODIST test version

The Enron Email Corpus was made public in the mid-2000s by the FERC.

Contents has been widely used by computer scientists to understand email behavior

Seen more limited usage in social scientific research (e.g., Aven, 2015; Benke 2018)





| How do users navigate the empty search box?

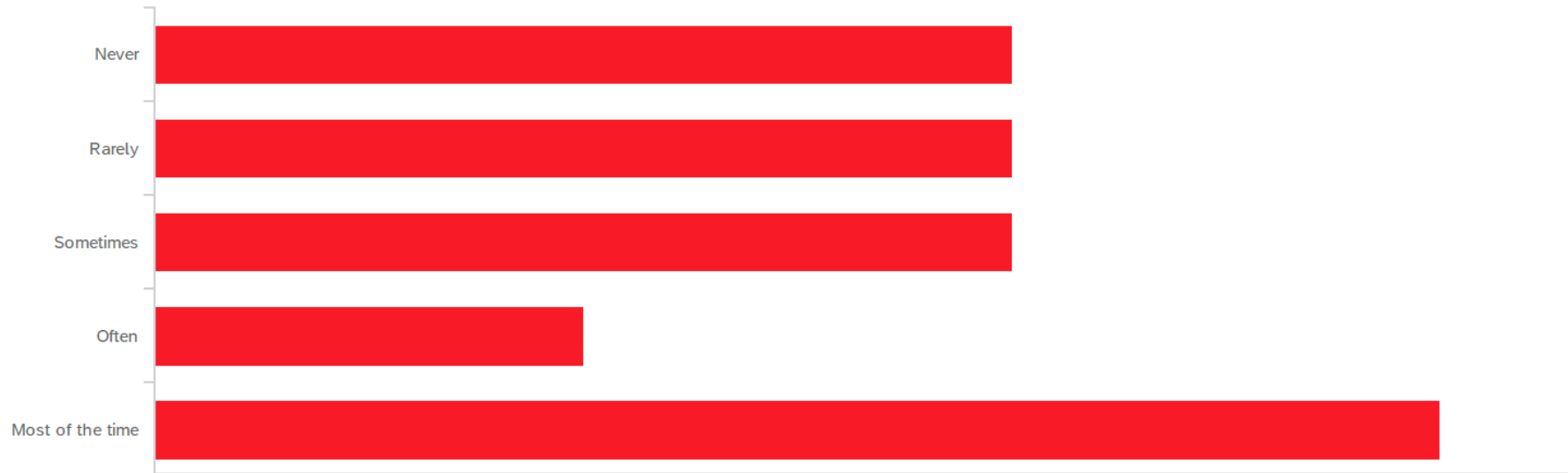
Experiments in Knowledge Discovery

Sample Searches

- Trading Strategies, Gaming, Manipulation
- Skilling, CEO, leaving, resignation
- Bankruptcy, failure, job losses
- Energy regulation, caps, orders,
- Pollution permits, green, generators
- Exporting, energy, California
- Crisis, political, energy

Initial Results:

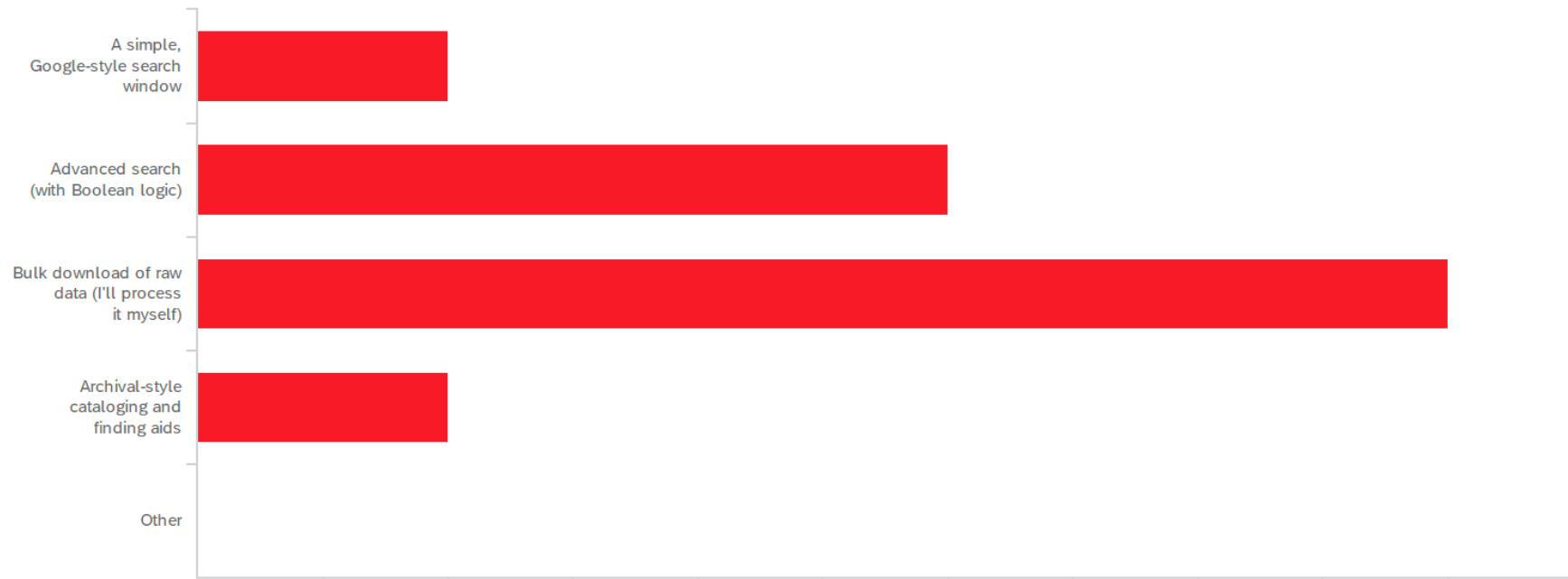
How often have you used born-digital data and/or resources as part of your research practice?



Initial Results:
In conjunction with what sort of methods do you use born digital data and resource?



Initial Results: What sort of access interface best suits your own research needs?



Challenges / Next Steps

- Computational intensity leads to high latency
- Build out “local” executable version (PC done; Mac coming)
- Engage more knowledge discovery-oriented users
- Conceptual Pillars: Genre, Provenance & Systematicity



EMCODIST - Desktop

EMCODIST - **Desktop** is a Desktop application. Currently developed for Windows OS.

The Desktop version helps to perform context based search facility on sensitive Email collections. Simple instructions on how to install this software on your Windows desktops are provided in the following pages.

Thank you!

Questions and Comments Welcome!



David A. Kirsch

dkirsch@umd.edu

