INTERNET ARCHIVE

Challenges of Collecting & Preserving the Social Web

- Not all data can be crawled
- Research and experimentation are essential to keep pace with publisher innovation
- Publishers can choose to “opt in” or “opt out”
- Harvested data is hard to make accessible
- Only a fraction of the social content on the Web is visible to anonymous visitors
- Almost all views are personalized

Primary Approaches to Social Media Capture Today

Web Crawlers (e.g. Heritrix)

Still a robust solution able to capture most content, including JavaScript. Often best method for capturing embedded media or media from accounts you don’t “own”/operate.

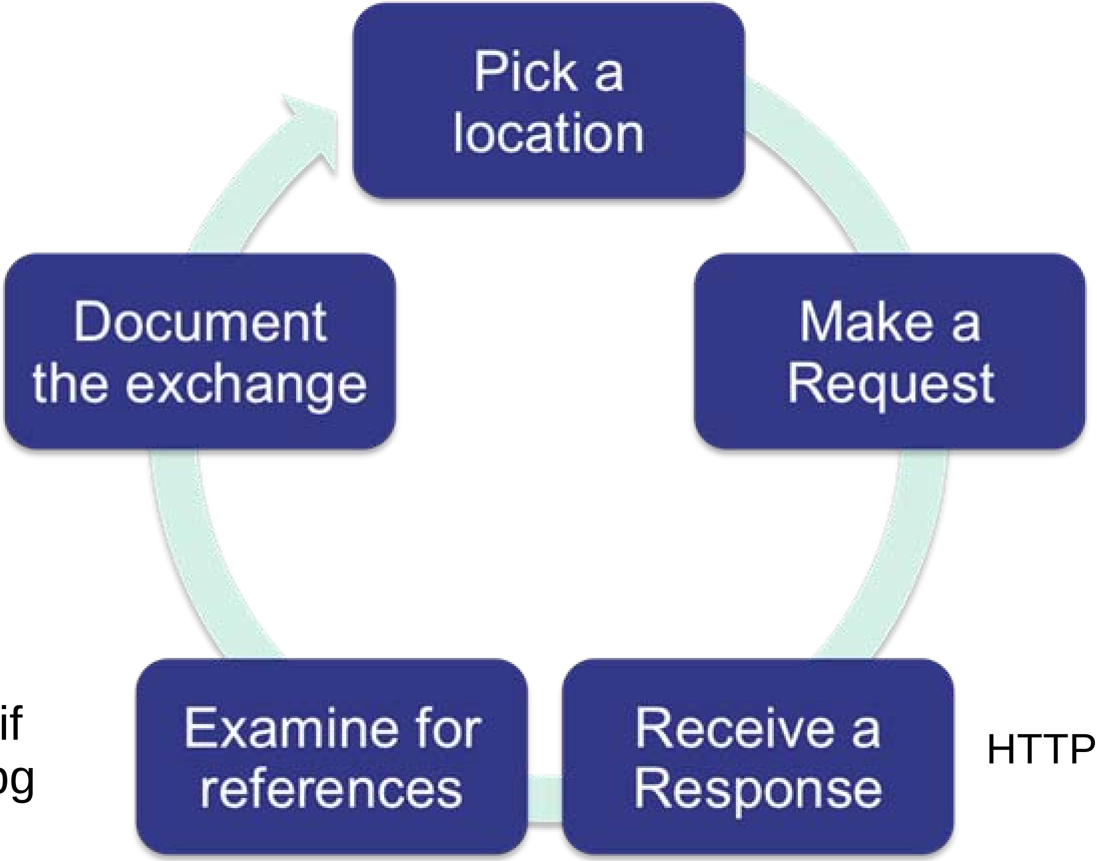
Scripted Browsers/Browser simulations

Rather than an automated crawl, a browser based tool can be instructed to mimic the behavior of a user and to archive what the browser “sees”.

APIs

Subscription/account based server request/responses that often include metadata not available to a crawler or browser. Often used to archive social media accounts you “own”/operate.

Classical Harvest Model: Crawling



Pull from queue

Pick a location

Make a Request

GET /

Receive a Response

HTTP/1.0 200 OK

Examine for references

text/css
image/gif
image/jpg
video
JavaScript

Document the exchange

W/ARC -

Differences Between a Crawler and a Browser

- Browsers grab all embedded resources as soon as possible
  - Typical behavior is a burst of traffic followed by long pauses.
- Crawlers have to play by different rules
  - Typical behavior is sustained traffic.
    - Can quickly overwhelm a website
    - Must apply intentional delays
  - Must obey robots.txt rules

Ongoing Experiments & Implementations

Open Planets (browser extractor module as alt for link extractor in H3):

- **https://www.github.com/openplanets/wap**

INA/IIPC (browser w/inline caching proxy; simulates user actions, outputs to WARC):

- **https://github.com/davidrapin/fantomas**

NDIIPP/NDSA (integrated crawler & browser w/caching proxy…):

- **https://github.com/adam-miller/ExternalBrowserExtractorHTML**
- **https://github.com/adam-miller/phantomBrowserExtractor (PhantomJS behavior scripts)**

Merging Browsing & Crawling: How Much is Gained?

Traditional Link Extraction: Baseline Test

  - 7444 URIs (200 response)
  - 795 URIs (404 response

- **Browser only** (full instance or scripted headless)**: ~30% less content**
- **PhantomJS** (WITH traditional link extractor)**: +24%**

  + Significant improvement in unique URI detection

  - Additional processing overhead

    …but can distribute load to dedicated browser nodes

  + Browser downloads in a separate workflow, asynchronous from Heritrix

  + JavaScript analytics

Other Strategies & Implementations

Data Mining & Analytics

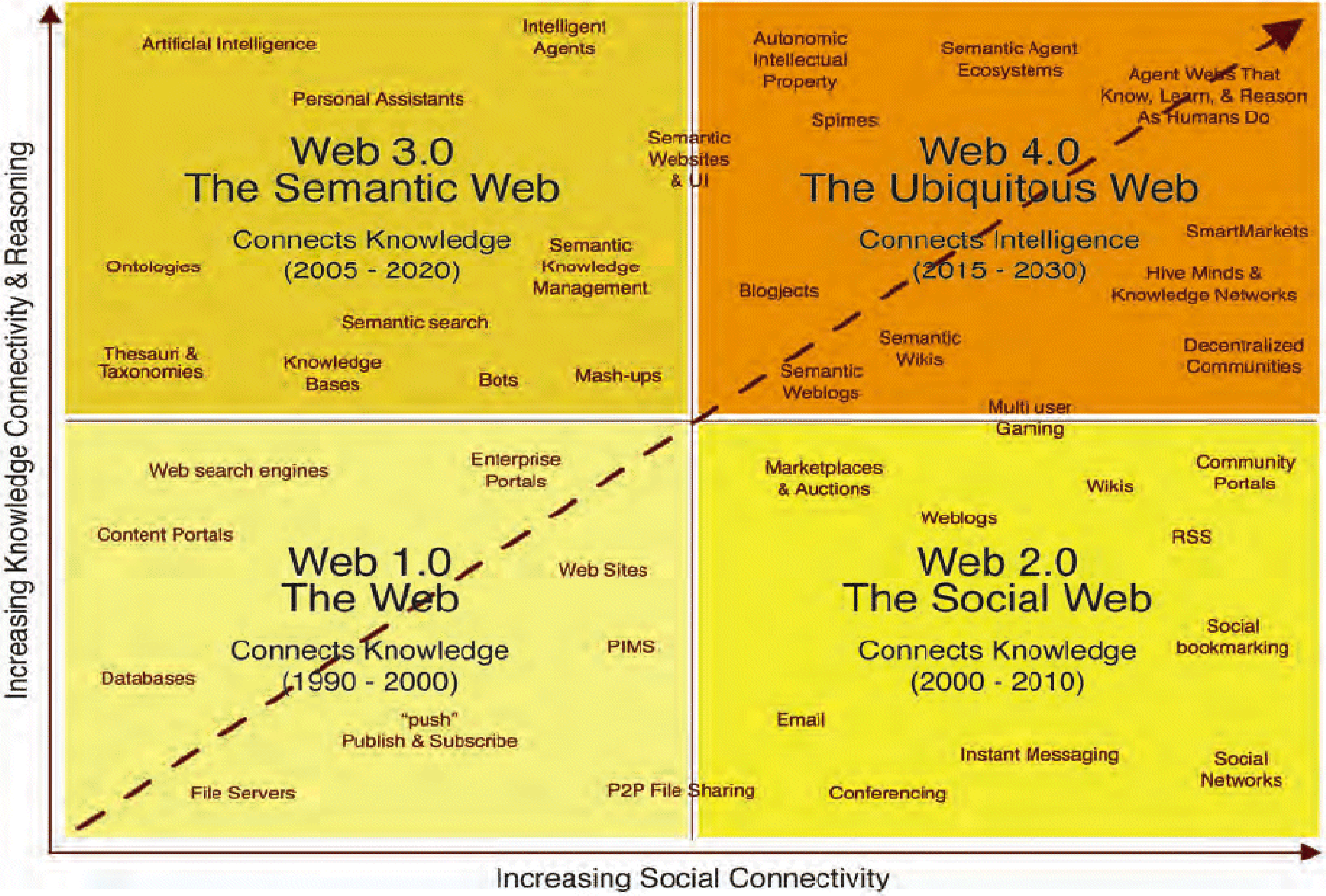
- Pre-Crawl Seed & Link Analysis
- Link/Script Analysis during an Active Crawl
- Post Crawl Link/Script Analysis, Patching & Auto QA

Native Feeds, APIs & Alternate Capture Methods

- Data format and context is as important as the content
- E.g. ArchiveSocial

Snapshot Generation & Recording

Internet Evolution



Artificial Intelligence

Intelligent Agents

Personal Assistants

Web 3.0
The Semantic Web

Connects Knowledge
(2005 - 2020)

Ontologies

Semantic Knowledge Management

Semantic search

Thesauri & Taxonomies

Knowledge Bases

Bots

Mash-ups

Semantic Websites & UI

Autonomic Intellectual Property

Semantic Agent Ecosystems

Agent Webs That Know, Learn, & Reason As Humans Do

Spimes

Web 4.0
The Ubiquitous Web

Connects Intelligence
(2015 - 2030)

SmartMarkets

Blogjects

Hive Minds & Knowledge Networks

Semantic Wikis

Decentralized Communities

Semantic Weblogs

Multi user Gaming

Web search engines

Enterprise Portals

Content Portals

Web 1.0
The Web

Web Sites

Connects Knowledge
(1990 - 2000)

PIMS

Databases

"push"
Publish & Subscribe

File Servers

P2P File Sharing

Marketplaces & Auctions

Wikis

Community Portals

Weblogs

RSS

Web 2.0
The Social Web

Connects Knowledge
(2000 - 2010)

Social bookmarking

Email

Instant Messaging

Social Networks

Conferencing

Increasing Knowledge Connectivity & Reasoning

Increasing Social Connectivity

Source: Nova Spivak, Radar Networks & Mills Davis, Project10x