Ethics – A Disclaimer

I am **not a legal expert**

Legislation changes with time and jurisdictions

All researchers must adhere to the ethical standards set out by their **Ethical Review Boards**

What Does Social Media Cover?

Social Media; My necessarily vague definition: Internet services to which users contribute **information**

Information

Open Text
The written word

Multimedia
Images, audio, movies

Other User Interactions
Mouse movements
Scrolling
Icon clicks

Primary Internet Services

Social Networking Services
Twitter, Facebook etc.

Modern Multimedia Services
YouTube, Twitch etc.

“Legacy” Services
Forums, blogs etc.

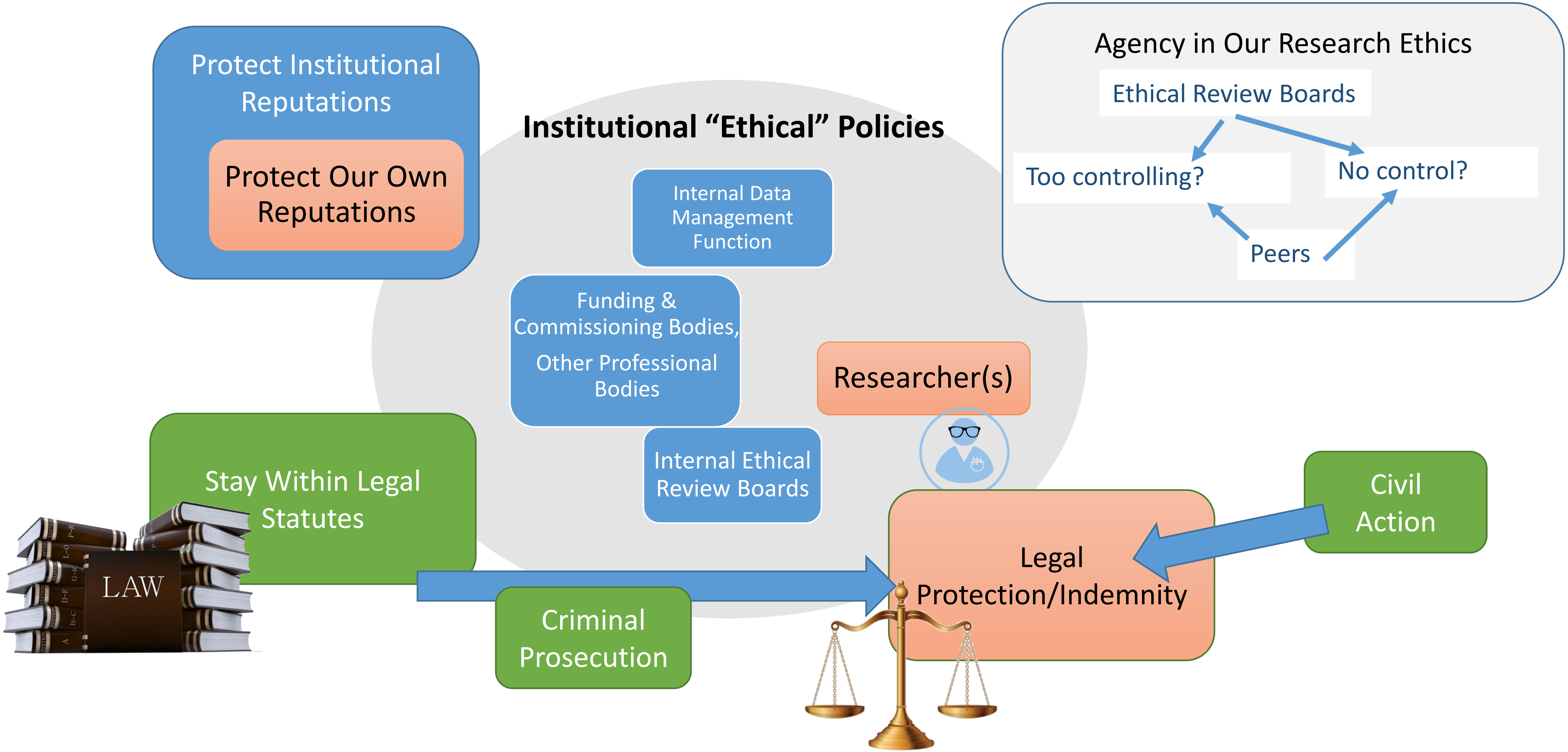
Other Legacy Services
Usenet, IRC etc.

Aggregation Services

Google Trends

Many Others ....

Amoral Reasons We Should Care



Protect Institutional Reputations

Protect Our Own Reputations

Institutional “Ethical” Policies

Internal Data Management Function

Funding & Commissioning Bodies,

Other Professional Bodies

Internal Ethical Review Boards

Researcher(s)

Agency in Our Research Ethics

Ethical Review Boards

Too controlling?

No control?

Peers

Stay Within Legal Statutes

Criminal Prosecution

Legal Protection/Indemnity

Civil Action

Relevant Legislation – Copyright and Data Privacy

	© Copyright	Data Privacy
Issue	• Copyright stops others using literary and non-literary work **without permission**. In the UK it **explicitly includes web-content** and even **databases**. • **Scraping data creates a copy** of that data. • Publication of raw data in research outputs. • Site Ts & Cs may explicitly state no use or copying of content.	• Protects individuals in law from the unauthorised retention and publishing of their **personal data**. • Pseudo-anonymous data is likely to still be protected • Processing of Special personally identifiable information is explicitly prohibited without explicit consent.
Position	• **Non-commercial research** is **protected under “fair-use”** or “fair-dealing”. • Berne convention grants **copyright exemptions where “the interests of right holders are not prejudiced”**. • UK Law provides **specific exemptions for data mining** in scientific research.	• All information should be anonymised, removing any data privacy concerns and risk of harm. • **Social media and search engines can make it trivial to identify an individual.** • Even if the data is public, raw data should be handled with the expectation it contains personally identifiable information.
But	• Data **must be lawfully accessed**	• Truly anonymous data is not protected • Must “relate” to the individual, not merely identify them • There are derogations under EU law for handling PII for research purposes.

Relevant Legislation - Human Subjects Research

Human Subjects Research is controlled under legal statute in the UK

Human Subjects Research Definition (WHO)

> *"any … systematic collection or analysis of data … to generate knowledge, in which humans are:*
>
> *i) exposed to manipulation, intervention, observation, or* ***other interaction with investigators directly*** *or through alteration of their environment, or*
>
> ***ii) become individually identifiable*** *through investigator's collection, preparation, or use of biological material or medical or other records"*

Ethical Tenets in Human Subjects Research - The Declaration of Helsinki

i) Do the most good
ii) Do no harm
iii) Respect for person, who exercise choice in participation through their **informed consent**
iv) And Justice, with a fair distribution of risk and benefit across participants

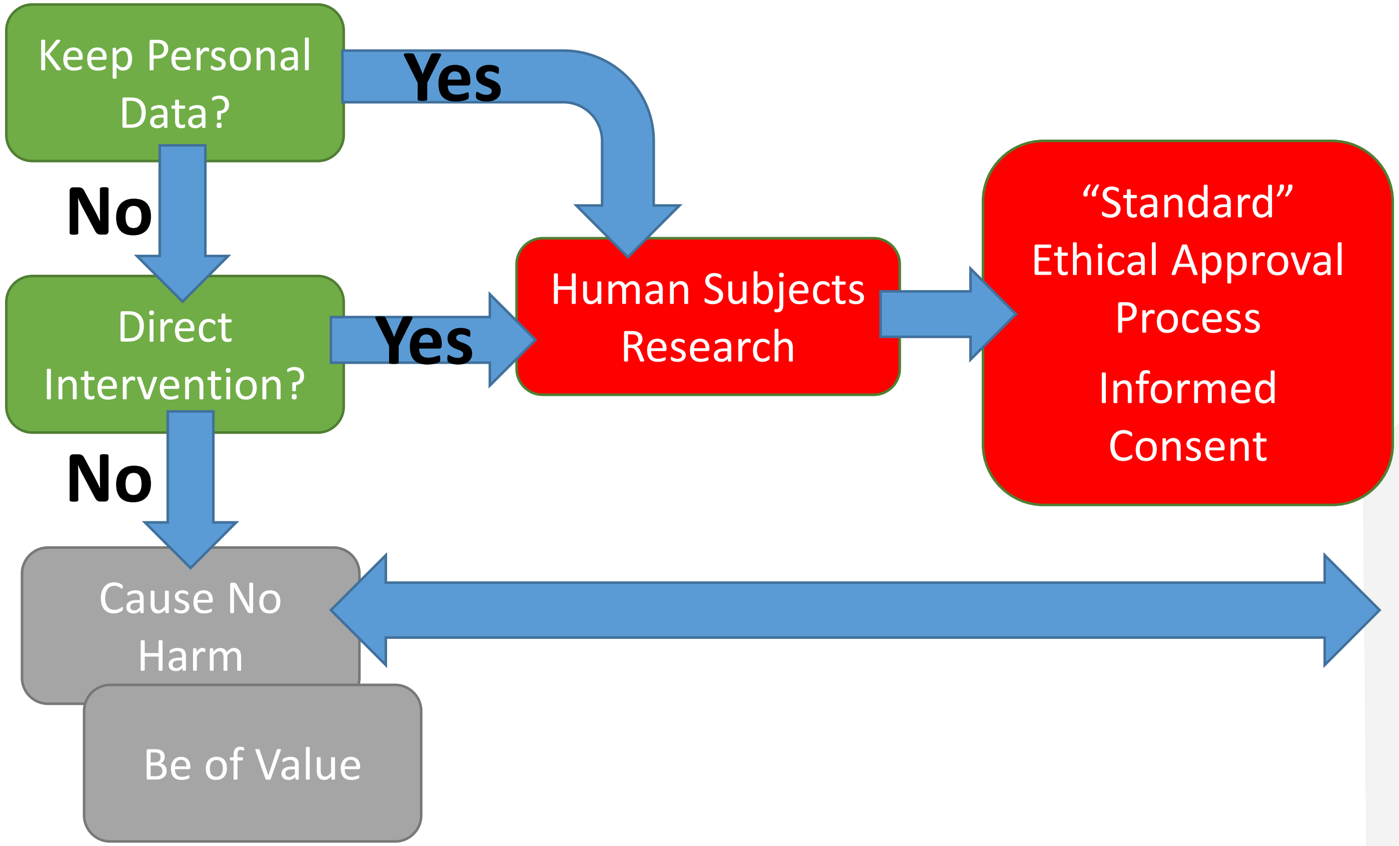
Hence, if our research involves human subjects, we need informed consent from all participants
BUT
Obtaining informed consent is on primary internet services
and informed consent includes the **right to withdraw** ones information and be **debriefed**

Social Media Research Types and Human Subjects Research

	Observational	Interactive	Survey Like
Researcher Intervention	✗	?	✓
Description	Data mining and analysis of public service user content with no service-enabled barriers to access. De-facto public, anonymity variable	The researcher is required to interact through the primary service to obtain data De-facto public, anonymity variable	"Traditional" surveys, with methodological elements supported by social media services. Private data, anonymity controlled
Example(s)	* Data mining of public forum posts. * Analyses of website hits.	* Users responding to researcher published tweets. * Researcher accessing posts to a private Facebook group	* Users recruited via social media to complete an online survey.
Informed Consent	✗ **Impractical**	? **Depends**	✓
Is Human Subjects Research?	**No**, provided **anonymity is not compromised**.	Requires careful consideration of the "interaction".	**Yes**, follows standard ethical processes for human subjects research.

Legalities Aside - Research Ethics

In Summary:


Keep Personal Data?

Yes

No

Direct Intervention?

Yes

No

Human Subjects Research

"Standard" Ethical Approval Process

Informed Consent

Cause No Harm

Be of Value

Stakeholder Groups

Social networking service businesses

Other Researchers

"Participants" and humans for which the ecological resources have a socioeconomic value

Image sources: wikimedia commons

Do No Harm to Other Stakeholders

Service Businesses

Considerations

- Reputational damage arising from the disparity between users' illusory perception of operating in a private space.
- Site terms and conditions of use, not legal documents but following these reduces potential for reputational damage.
- Scraping activity may have negative impacts on services.

Scraping Etiquette

- Do not make excessive data requests.
- Execute scraping during times when site traffic is at a minimum.
- Use the services API, or a wrapper for the services API if available.
- Consider using Google's website cache, or archival services such as archive.org.
- Respect the service's robots.txt and robots meta tags in the root of a
- **Do not circumvent technical measures (lawful access!)** which limit or prevent content scraping or circumvent access restrictions (e.g. download content from private groups).
- Do not attempt to mask your IP address by using proxy services. Ensure request headers include contact details.

Other Researchers

Considerations

- Inappropriate use of social media data, or ignoring T&Cs may lead to loss of access.
- Negative reactions by online communities may cause mistrust of researchers and stymie other avenues of research, e.g. engagement with citizen science projects.
- Communities may fail to differentiate between the researcher and other organisations. This may lead to mistrust of other organisations which were not directly involved in the research.

Social Media Users, Communities with Real-World Investment

Considerations

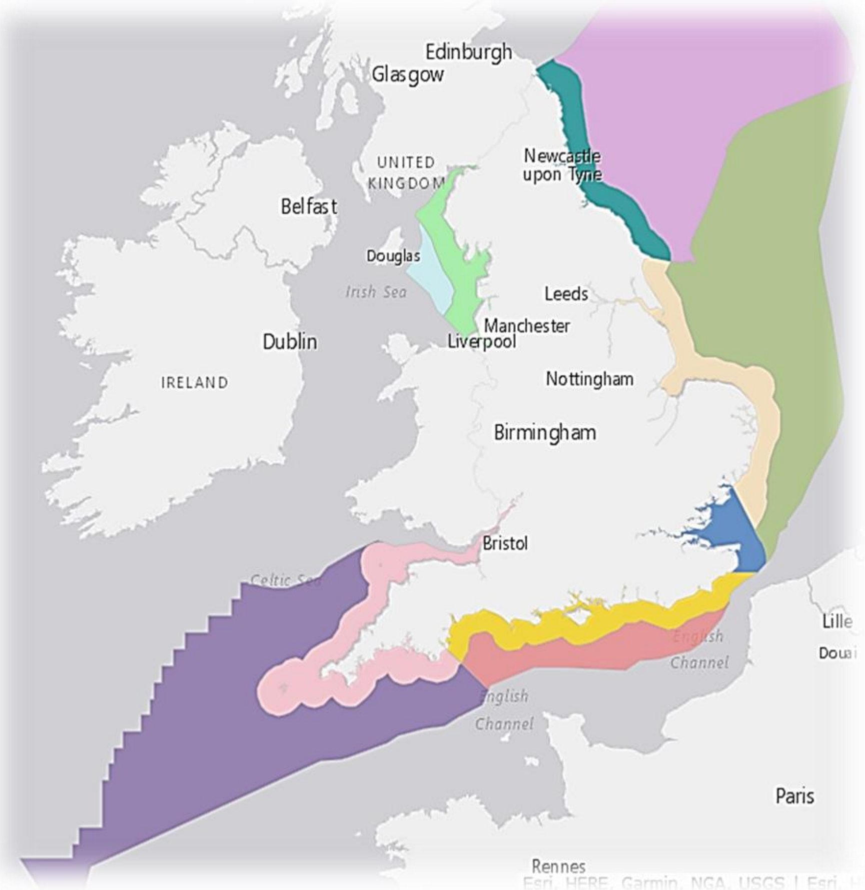
- Any harm arising from de-anonymised social media users.
- The effect of research outputs on management policy. Research can lack transparency.
- Lack of engagement with stakeholders may increase risk of non-compliance with regulations or other management actions.
- Information on ecological resources could change human behaviour (e.g. fishing practices) and affect species and their environment.

Sea Angler Mapping for the MMO

Aim

- To provide high resolution maps on the spatial and temporal distribution of sea angling "effort".
  - By season
  - By species (cod, bass etc)
  - By platform (shore, charter boat etc.)

England's Marine Planning Areas



What did I do?

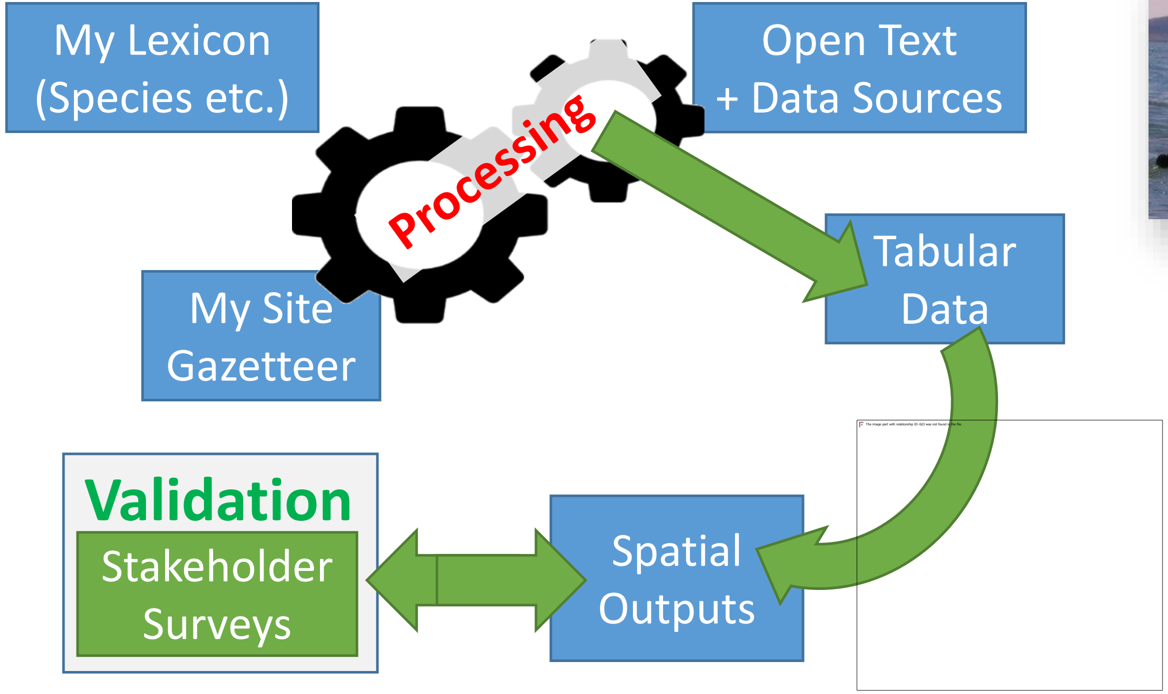
- Mined open text data from angler-centric social media sites (>90% forums)
- Produced a **qualitative** indicator of "effort" by **Species** *x* **Season** *x* **Platform** for shore angling
- Validation by Survey

Marine Management Organisation

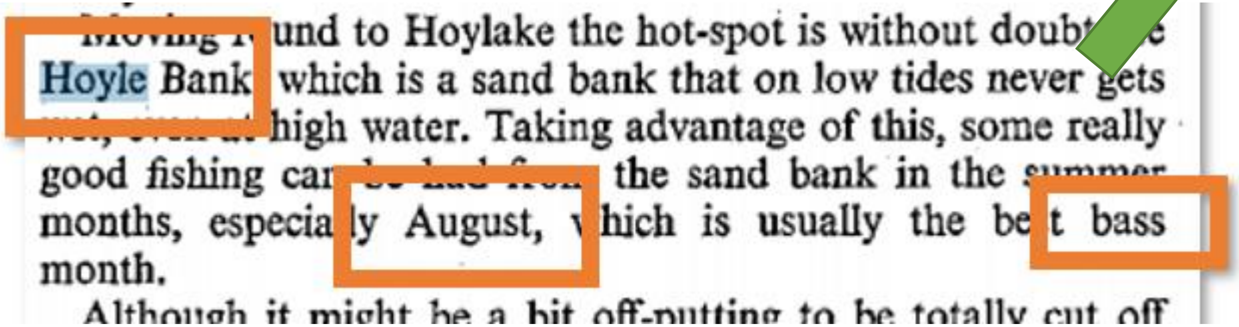
Cefas

substance.

MMO – Inputs and Outputs



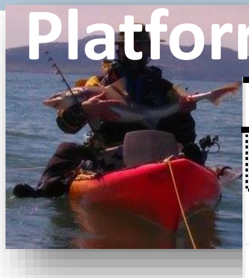
With the lexicon and gazetteer open text is turned into …….

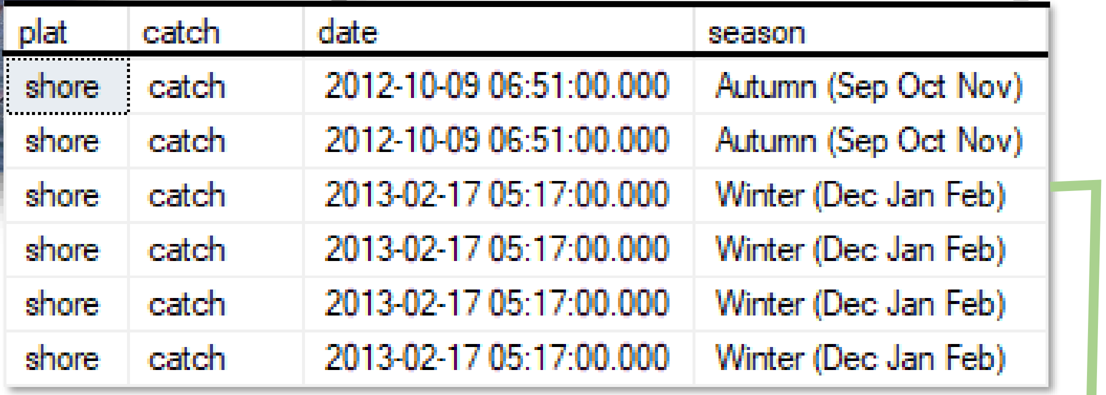


Moving round to Hoylake the hot-spot is without doubt the Hoyle Bank which is a sand bank that on low tides never gets wet, even at high water. Taking advantage of this, some really good fishing can be had from the sand bank in the summer months, especially August, which is usually the best bass month.

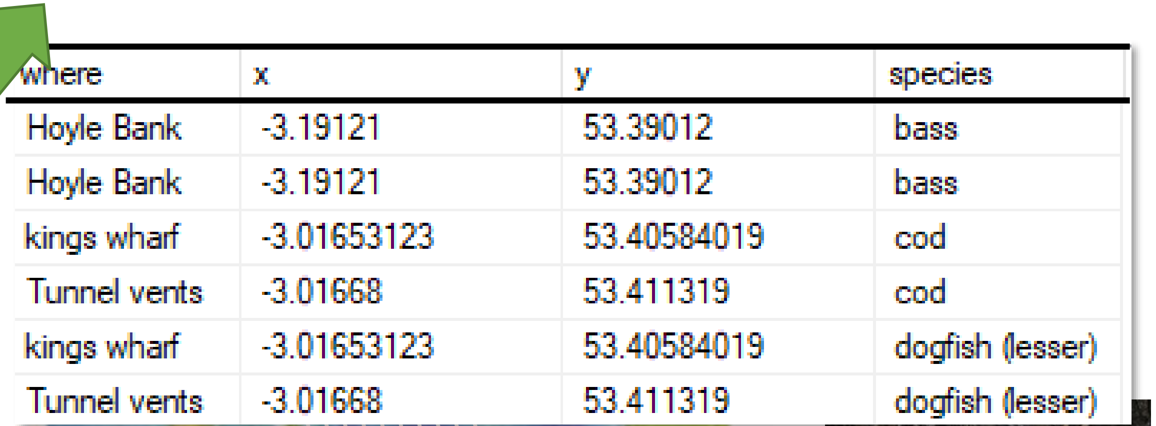
Although it might be a bit off-putting to be totally cut off

…… tabulated data





plat	catch	date	season
shore	catch	2012-10-09 06:51:00.000	Autumn (Sep Oct Nov)
shore	catch	2012-10-09 06:51:00.000	Autumn (Sep Oct Nov)
shore	catch	2013-02-17 05:17:00.000	Winter (Dec Jan Feb)
shore	catch	2013-02-17 05:17:00.000	Winter (Dec Jan Feb)
shore	catch	2013-02-17 05:17:00.000	Winter (Dec Jan Feb)
shore	catch	2013-02-17 05:17:00.000	Winter (Dec Jan Feb)

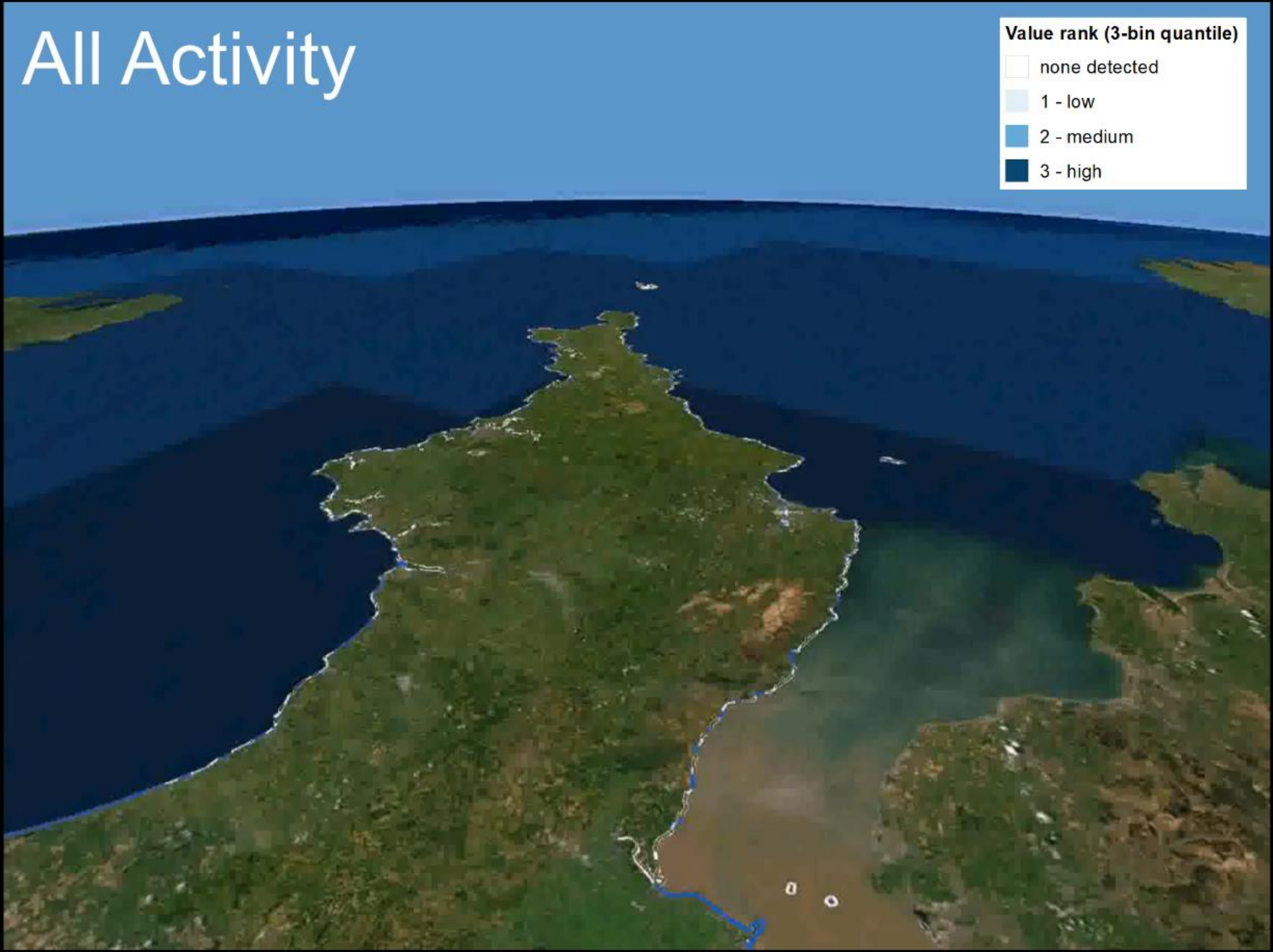


where	x	y	species
Hoyle Bank	-3.19121	53.39012	bass
Hoyle Bank	-3.19121	53.39012	bass
kings wharf	-3.01653123	53.40584019	cod
Tunnel vents	-3.01668	53.411319	cod
kings wharf	-3.01653123	53.40584019	dogfish (lesser)
Tunnel vents	-3.01668	53.411319	dogfish (lesser)

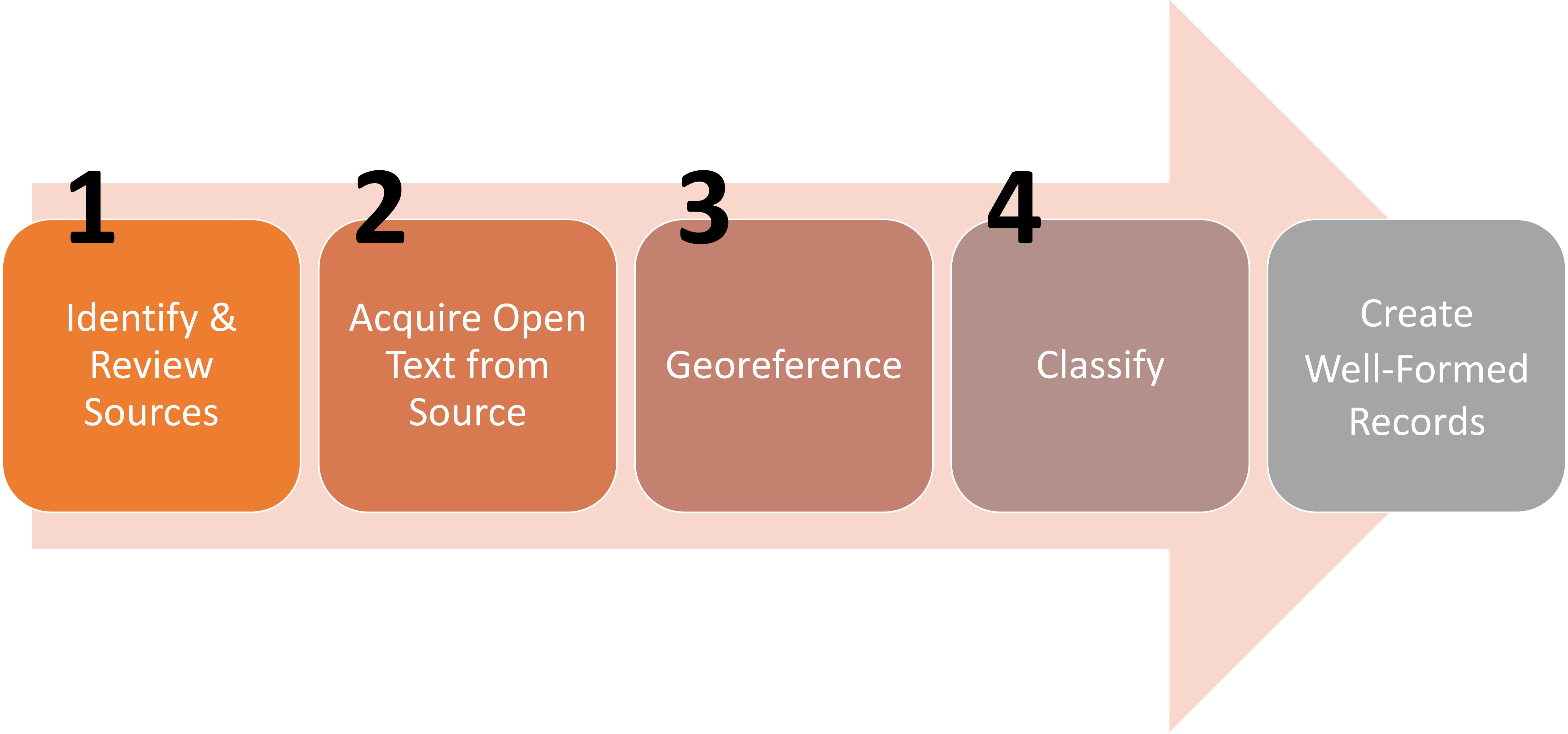
Species

Image sources: the author, wikimedia commons, Fishing in the North West by B. Gledhill

All Activity



MMO Text and Data Mining – Process Overview



1 Identify & Review Sources

2 Acquire Open Text from Source

3 Georeference

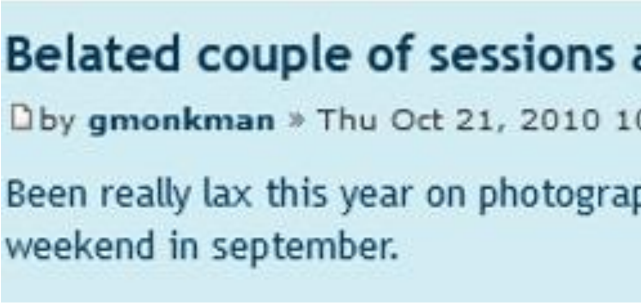
4 Classify

Create Well-Formed Records

Sources

- **Reviewed potential sources**
  - Sea Angling Magazines
  - Forums
  - Blogs and Static Websites
  - Surveys and Technical Reports
  - Angler Volunteers

- **531 Data Sources Reviewed**
- **477 (90%) were from fisher knowledge**



Belated couple of sessions at U

by **gmonkman** » Thu Oct 21, 2010 10:38

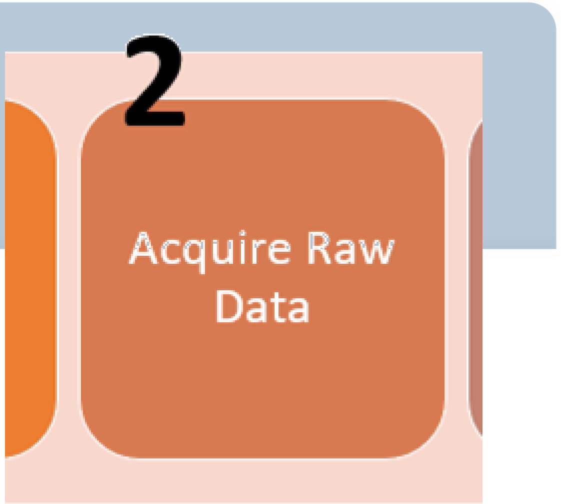
Been really lax this year on photography
weekend in september.

Today's best shore and boat angling magazine
SeaAngler
BIG ON BOAT ANGLING
Exclusive
WELSH CHARTER BOAT BREAKS NEW GROUND
LIFE TAG
BUOYS
UNIQUE
BASS ON BAIT AND LURES
THE BEST BEACH MULTIPLIERS TESTED
COD GALORE

[1] www.aquamaps.org, accessed 7/8/2015

Image sources: aquamaps.org, Bass Angler Mag., Google Maps, Heavy Metal Sea Danglers, Sea Angler Mag.

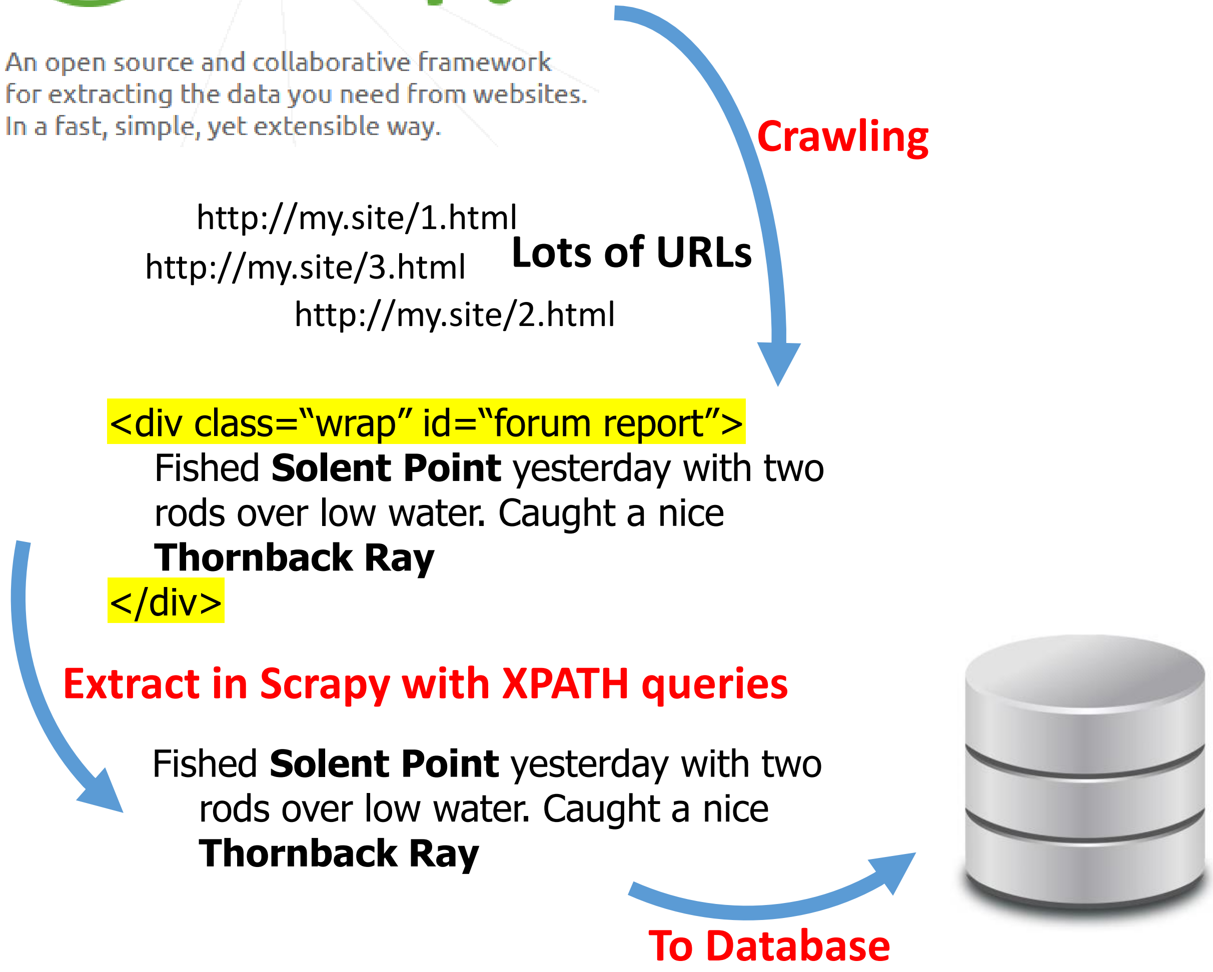
Acquire Raw Data



Scrapy

https://scrapy.org/

An open source and collaborative framework
for extracting the data you need from websites.
In a fast, simple, yet extensible way.



Crawling

http://my.site/1.html
http://my.site/3.html
http://my.site/2.html

Lots of URLs

```
<div class="wrap" id="forum report">
    Fished Solent Point yesterday with two
    rods over low water. Caught a nice
    Thornback Ray
</div>
```

Extract in Scrapy with XPATH queries

Fished **Solent Point** yesterday with two
rods over low water. Caught a nice
Thornback Ray

To Database

Data Scraped for This Project

Unique open text "samples"	~ 400,000
Word count	~ 35,000,000

Georeferencing

3

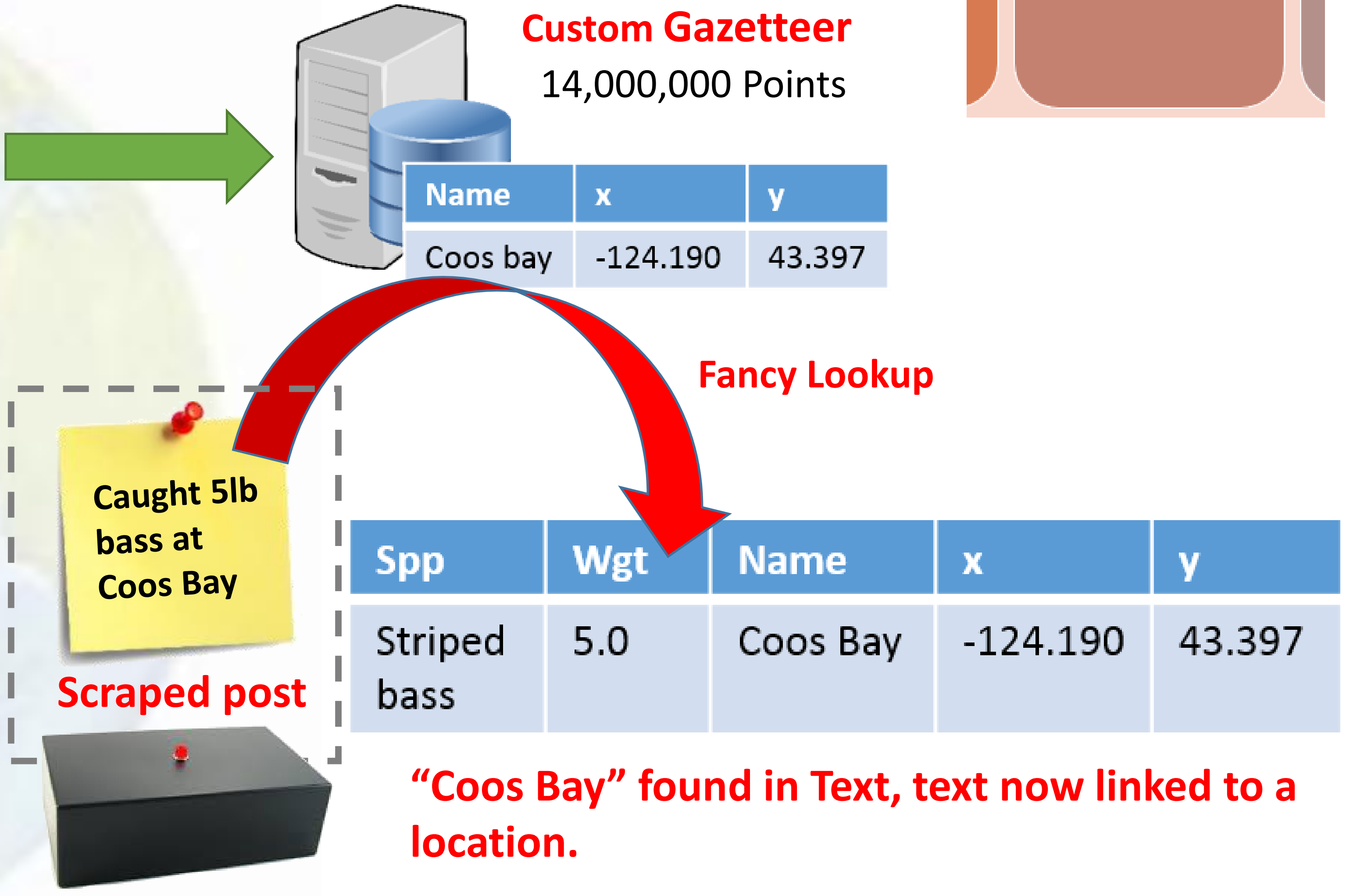
Georeference

- **Compiled Custom Gazetteer in a High Performance Database**

Source	Format
Volunteer markup in Google Earth; **Fisher Knowledge on Google/Bing Maps**	KML WGS84

geograph.org.uk;
geonames.org;
Ordnance Survey: Open Names, OS Locator;
Ordnance Survey
Others
	Excel, CSV
	OSGB36,
	WGS84

GPS devices UK Hydrographic Office	GPX WGS84
UKHO: Seacover_Polygons, shoreline_constructs, marine use, named sea features; **MEDIN**: sea features gazetteer	Shapefiles ETRS89, WGS84



Custom Gazetteer

14,000,000 Points

Name	x	y
Coos bay	-124.190	43.397

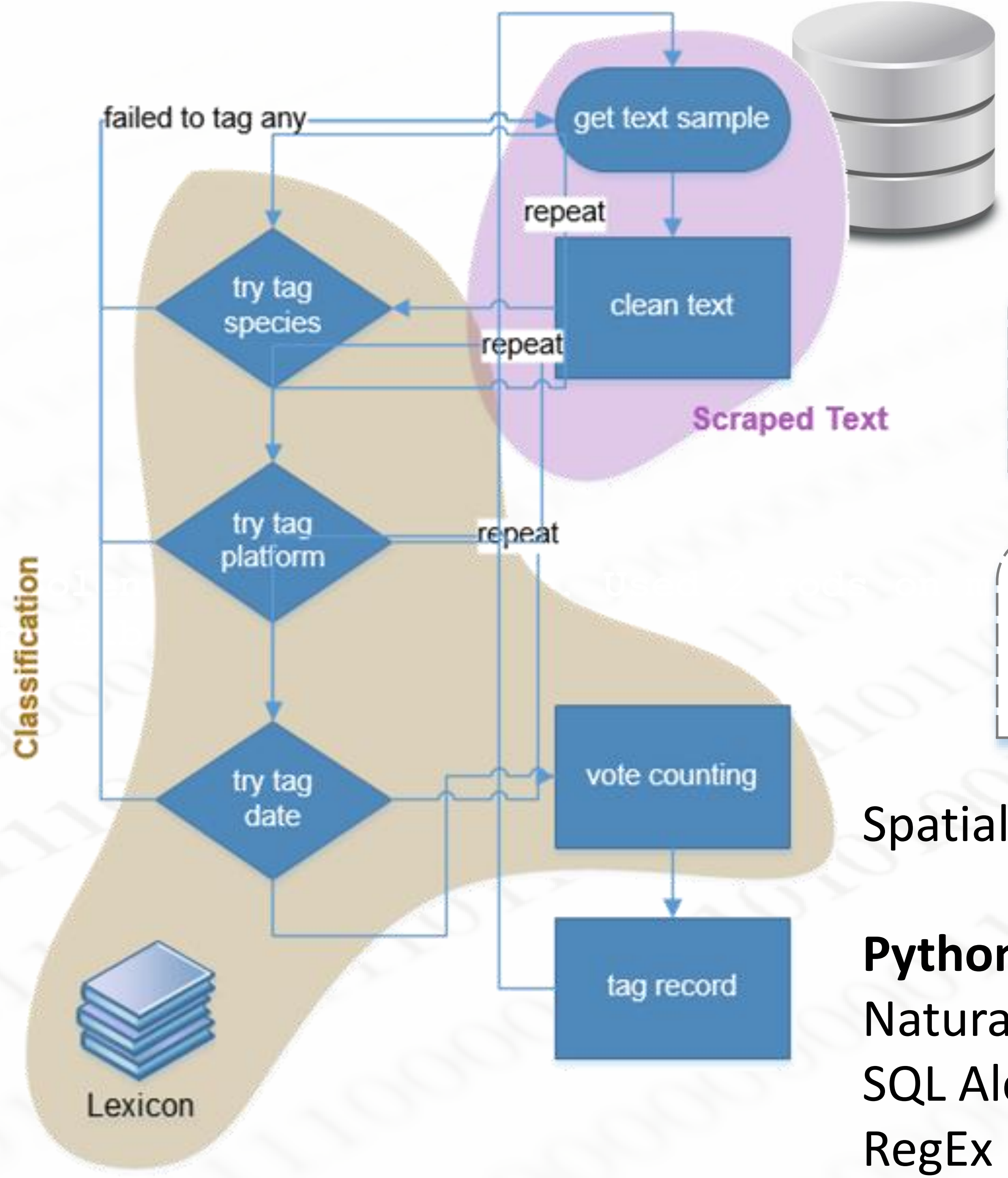
Fancy Lookup

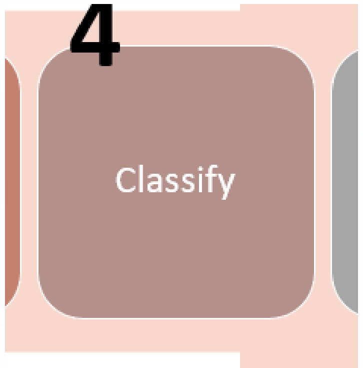
Scraped post

Spp	Wgt	Name	x	y
Striped bass	5.0	Coos Bay	-124.190	43.397

"Coos Bay" found in Text, text now linked to a location.

Image sources: Google Maps, Ordnance Survey, UKHO, Wikimedia commons





Species nouns
e.g. bass, cod

12

Gear nouns
e.g. rod, net

Platform related words
e.g. boat, Titanic, paddled, onboard

Nouns | Proper nouns | Verbs | Adjectives

Time
e.g. midnight, 12:45

Nouns | Time

Duration related words
e.g. ebb, arrived, before

Nouns | Verbs | Prepositions

Quantities
e.g. few, one, 2, 3.2

Quantifiers & determiners | Numerics

Spatial Queries in MS SQL Server

Python Packages

Natural Language Toolkit

SQL Alchemy

RegEx (text searches)

Last Slide – My Relevant Papers

Monkman et al. (2017). **The Ethics of Using Social Media in Fisheries Research**. *Reviews in Fisheries Science & Aquaculture.* https://www.tandfonline.com/doi/full/10.1080/23308249.2017.1389854

Monkman et al. (2018**). Heterogeneous public and local knowledge provides a qualitative indicator of coastal use by marine recreational fishers**. *Journal of Environmental Management.* https://doi.org/10.1016/j.jenvman.2018.08.062

Monkman et al. (2018). **Text and Data Mining of Social Media to Map Wildlife Recreation Activity**. *Biological Conservation.* https://doi.org/10.1080/23308249.2017.1389854

Monkman et al. (2020). **Mapping Sea Angling (MMO1163)**. Technical Report for the Marine Management Organisation. https://www.gov.uk/government/publications/mapping-sea-angling-mmo1163

gmonkman@mistymountains.biz