

Understanding the importance of Provenance from the Perspective of a (Geospatial) Decision-maker

Nikos Papapesios

Civil, Environmental and Geomatic Engineering
University College London

UCL Supervisors: Claire Ellul, Artemis Skarlatidou

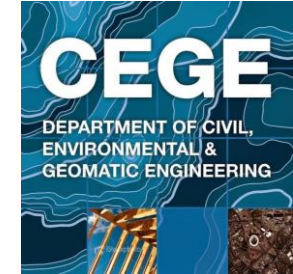
DSTL Supervisor: Amanda Shakir

Overview

- Research Focus
- Initial Research
 - Primary aim (Research sub-question)
 - Producer side investigation (standards and implementations)
 - Results (Theoretical Provenance Framework)
- Methodology 1
 - Approach decision-maker's perspective (Qualitative and quantitative data)
 - Qualitative data analysis (Thematic Analysis)
 - UpToDate Results
- Future Plans
 - Methodology 2

Research focus

- Civil, Environmental & Geomatic Engineering
- May 2017 – May 2020
- Co-sponsored
 - Defence Science and Technology Laboratory
 - UK Engineering and Physical Sciences Research Council
- 4 research stages (end-user perspective)
 - Provenance and related concepts investigation
 - Approach the most important provenance factors
 - Enhance the interaction between decision-makers and information outputs
 - Evaluate trust perceptions



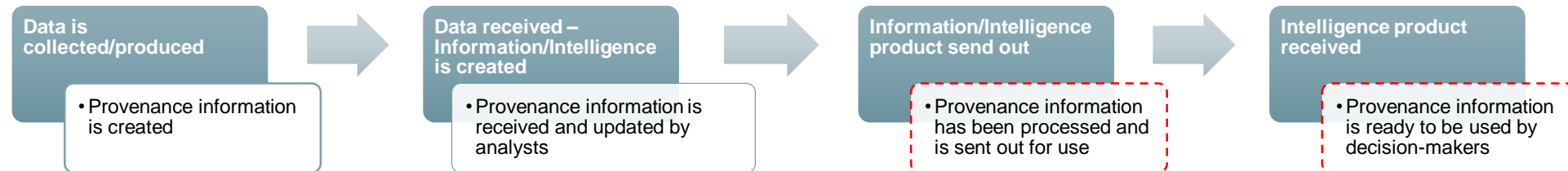
Research Focus

- **Can identifying and presenting the required provenance factors for information that has been derived from geospatial sources, in a usable and useful manner, help decision-makers to make use of this information?**
 - How provenance is linked with metadata and data quality?
 - What are the most important provenance factors according to decision-makers?
 - What is the best way to present them?
 - To what extent decision makers' trust perceptions can be influenced?

Initial Research – Linking provenance

- Provenance is defined as *“information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its **quality**, **reliability** or **trustworthiness**”* (W3C 2010).
 - Several provenance definitions, descriptions and characteristics are found
 - Data quality elements and indicators are examined
 - Metadata types and sub-metadata elements

Initial Research – Linking provenance

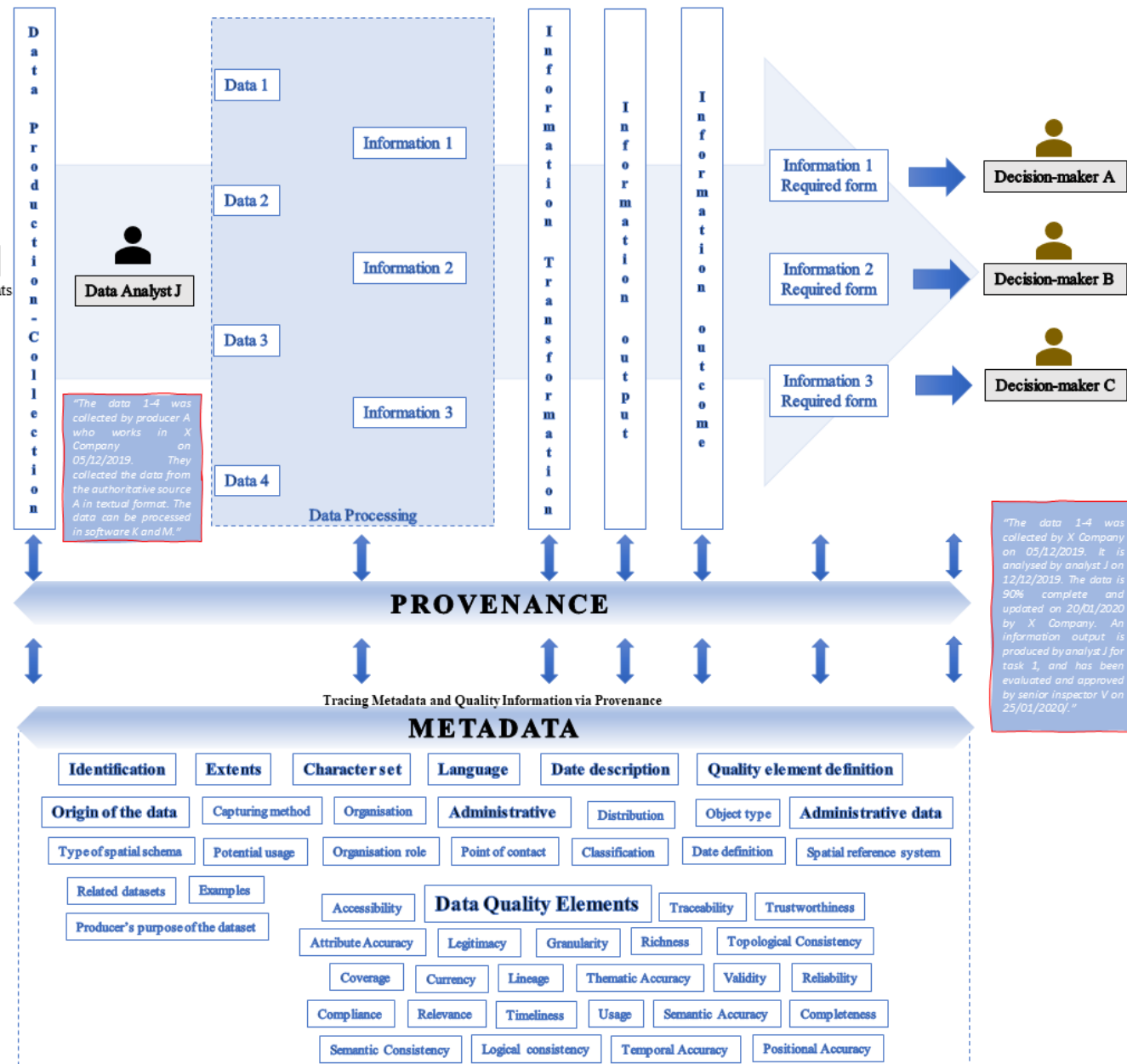
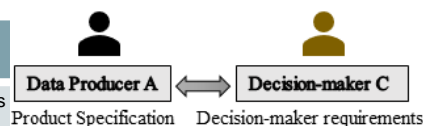


Initial Research – Linking provenance

Standardisation Body	Basic Description
Provenance International Organisation for Standardisation (ISO) • ISO 8000 – Part 120	Metadata It is a global alliance composed of 163 national standards bodies, having a goal to gather experts from all over the world and approach global challenges with innovative solutions (ISO 2018b). • ISO 19115-1
World Wide Web Consortium (W3C) • W3C PROV	It is an international organisation developing standards, protocols and guidelines to guarantee the usability of the web (W3C 2018). • STANAG 4774
Dublin Core Metadata Initiative (DCMI)	It is an open organisation focusing on metadata design and practices (DCMI 2018). • Dublin Core
Data Quality North Atlantic Treaty Organization STANdardization Agreement (STANAG)	Lineage Implementations Through standards, interoperability among NATO's allies is accomplished and by implementing several concepts, doctrines and procedure, the use of the available sources improves its effectiveness (NATO 2018).
Defence Geospatial Information Working Group (DGIWG) • ISO 8000	It is a multi-national body responsible for geospatial standardisation of defence organisations (DGIWG 2018). • OGC
European Committee for Standardization (CEN)/TC 287 • ISO 19157	It is the European Committee for Standardisation officially recognised by the European Union to develop and define voluntary standards (CEN 2018). It is composed of 34 National Members. • INSPIRE
Federal Geographic Data Committee (FGDC) • STANAG 2215 IGEO	The Federal Geographic Data Committee (FGDC) is a national (US) governmental committee, providing insight and oversight for geospatial decision-making (The Federal Geographic Data Committee 2018).

Initial R

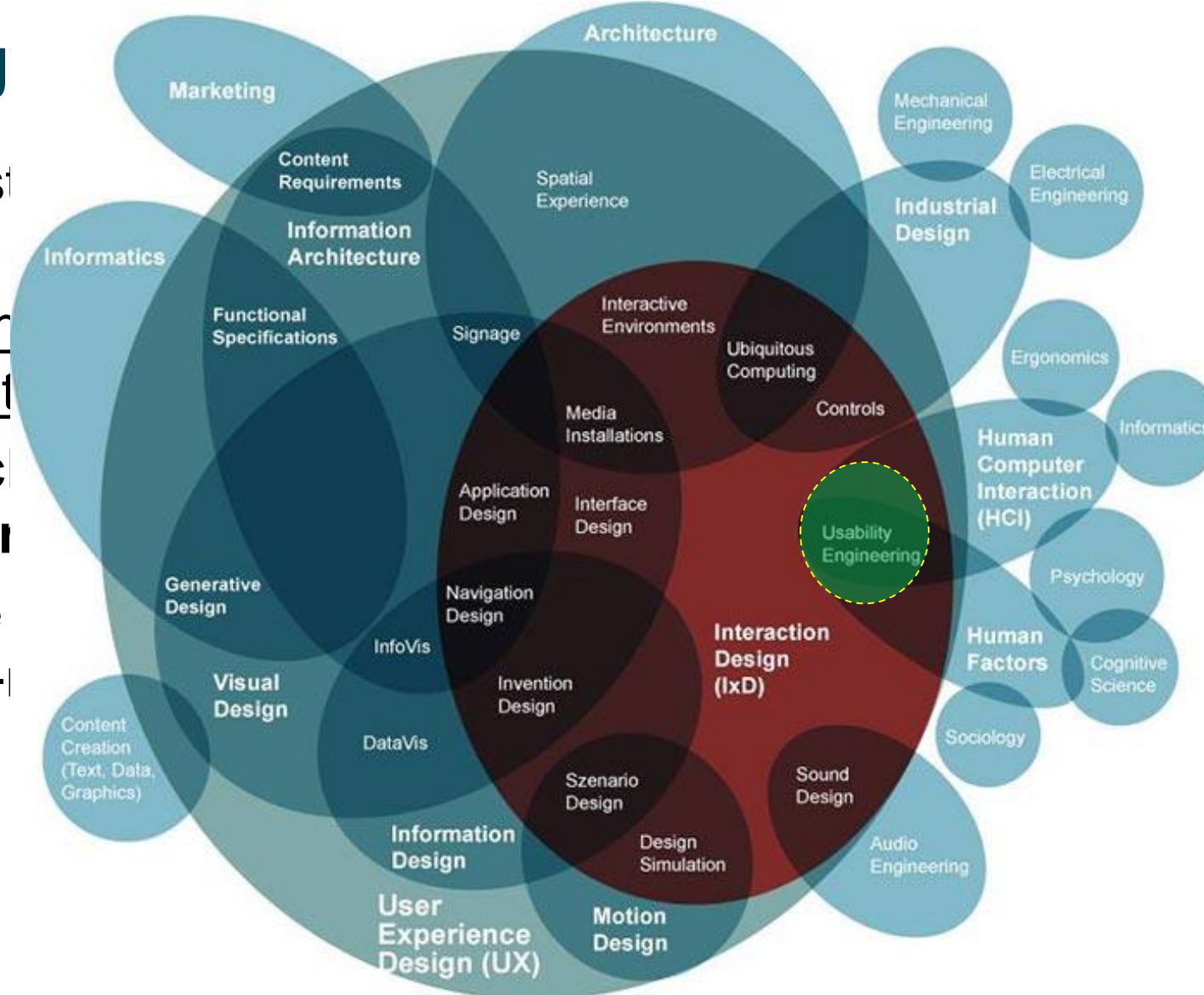
Org. Standards	
ISO	ISO is
8000	It is the
19157	It point concept
19115-1	It provide element
W3C	It is an web.
STANAG	Through
2215	It aims
4774	It provide
FCDC	It is a l
DCMI	It is an
Dublin Core Metadata Element Set	It provide
CEN/TC 287	It is re structu
DGIWG	It is a r



is and their application
9114 and 19113. It is
112, 19108, 19103,
standards.
39.85, and IETF RFC
OGC, and several
d. It follows INSPIRE
C 211, and OGC has

Methodology

- 4 research streams (from different perspectives)
- Provenance investigation
- **Approach** provenance
- Enhance decision-making outputs
- Evaluate







of a system for a McDonald and

f system's user valuation

Methodology 1

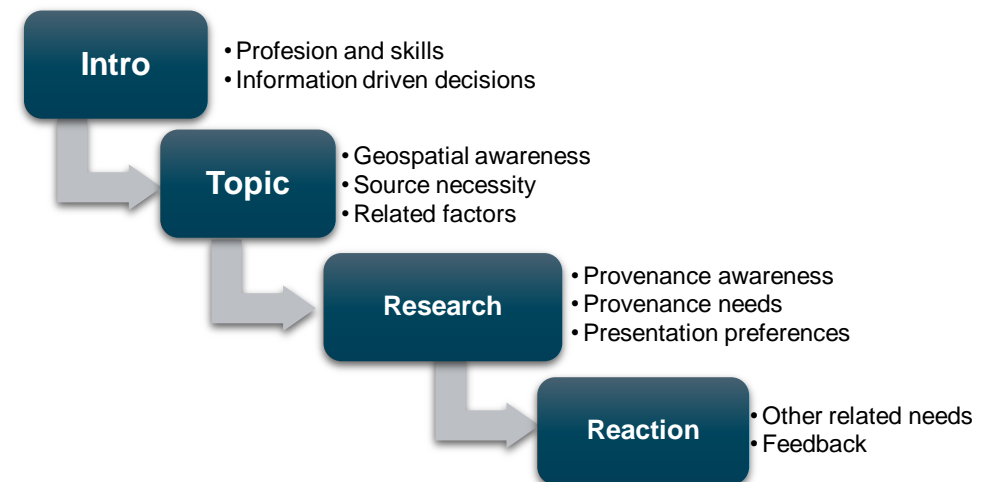
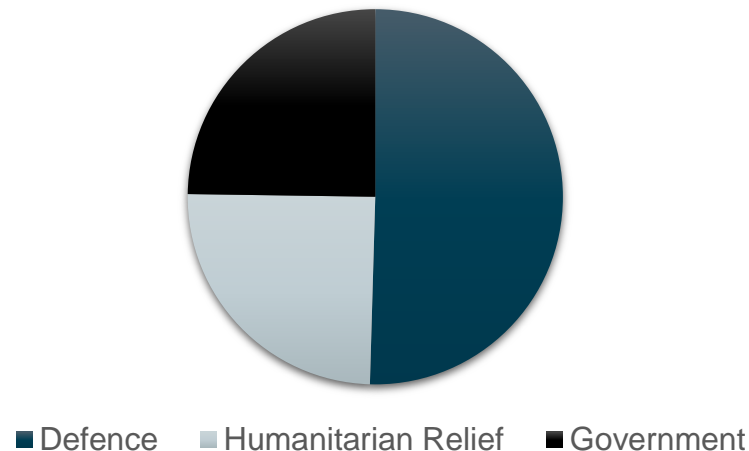


- Goal Setting: Identify the most important factors
- Provenance Requirements:
 - At least basic geospatial knowledge
 - Use information in any data stage before decision-making

Case Studies	Description of participant identification
Defence (RSMS) 	A snowballing method is used to identify participants in the UK defence, making use of the Defence Security Technology Laboratory (DSTL) network. Thus, a wider audience also containing participants from the Royal School of Military Survey (RSMS) accept to participate.
NGO (MSF) 	Three humanitarian NGOs which are in close cooperation accept to participate: Missing Maps, British Red Cross (BRC) and Doctors Without Borders (MSF).
GOVERNMENT (MOJ, Hackney)  	The GIS offices of the HM Courts and Tribunal Service (HMCTS) and of Hackney's municipality accept to participate as the governmental (central and local) case study, aiming to support decision-makers and to improve the spatial information of the municipality.

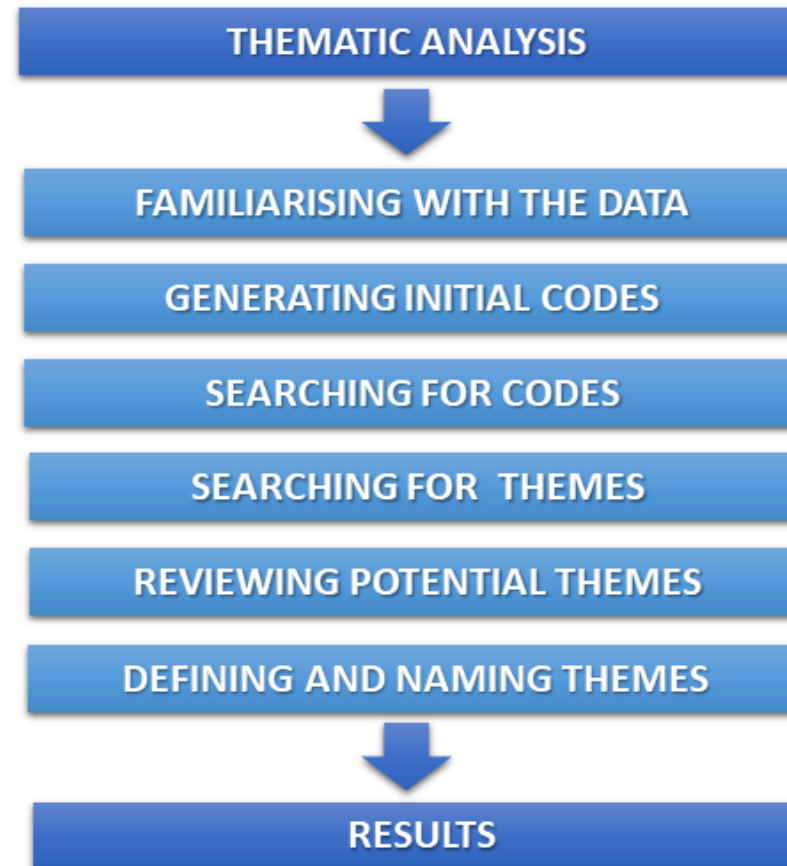
Methodology 1 – Approach Decision maker's perspective

Sectors Participation

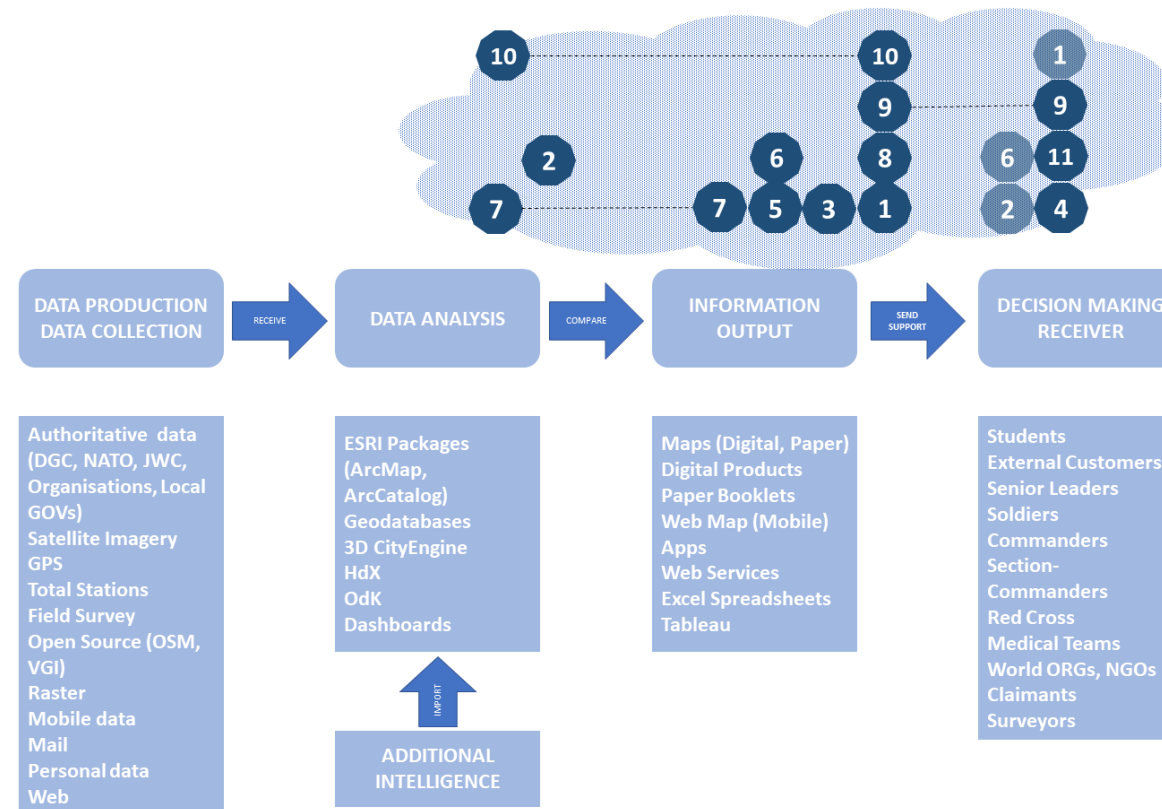


- 11 semi-structured interviews
- ~ 45 minutes each
- Full detailed transcriptions (121 pages, ~ 61K words)

Methodology 1 – Thematic Analysis



Methodology 1 – Familiarising with the data

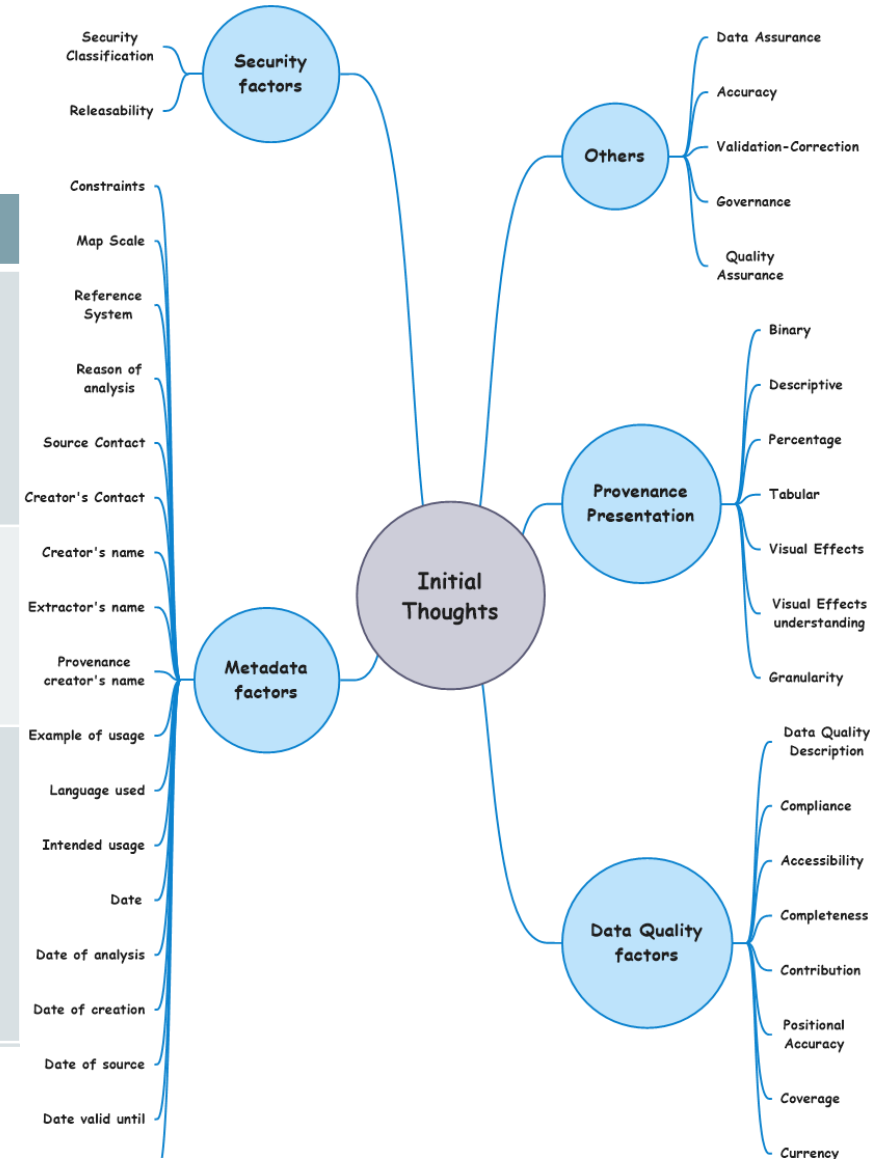


- The more the geospatial knowledge the less decision-making and vice-versa.

NVIVO

Methodology 1 – Generating initial codes and themes

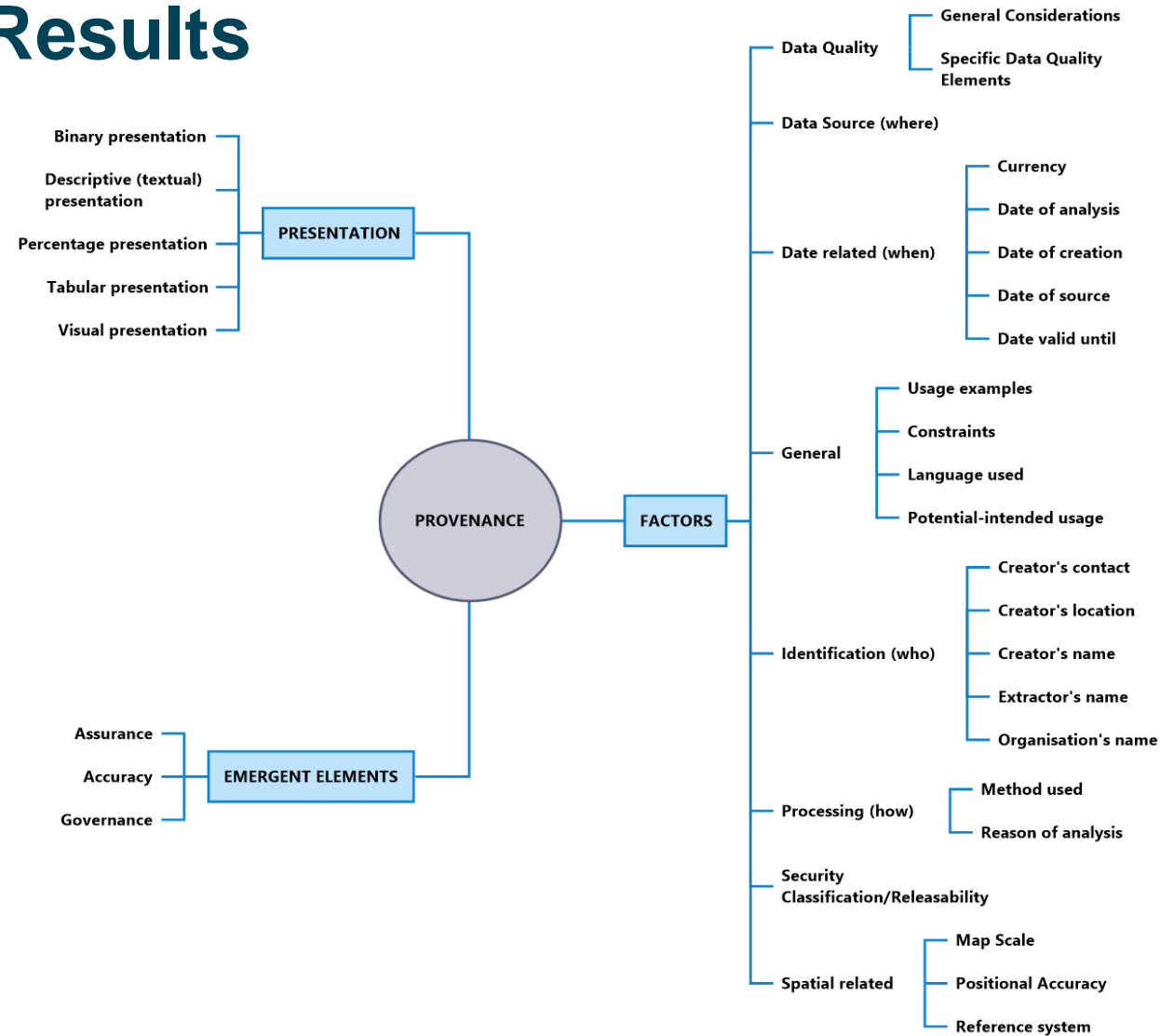
Transcript	Codes	Themes (sub)	Participant	Page
But like I said, when you know it's a product from a <u>specific organisation</u> , you know, <u>you already know majority of that information</u> . So, you know what <u>level it is collected</u> across that are here to a <u>standard</u> ; the bits of the information you would then <u>lose is how current is it, when was it collected?</u> Um, and stuff like that.	- Organisation name: Awareness of basic factors - Currency and date could be lost.	FACTORS Identification Date Related	3	7
For me, the <u>date goes also really on top</u> there. For me, <u>date is really a thing</u> . I would say if it's <u>date from the 1970s, okay</u> . (Smiling) <u>Just leave it out</u> . But that's true. You have those maps of the 1970s.	- Important of date - Currency important explained through example. - It seems that she can decide based on date	FACTORS Date Related	8	13
So, when I'm looking at this, it just <u>seems like it's all over the place</u> , whereas that's <u>very concise</u> . It's <u>easily readable and it's understandable</u> . Like I <u>understand</u> it <u>straight</u> , straight away. Um, so <u>if we can have the, the data provenance listed in the table, but also add a visual effect to go with it</u> . So <u>not only as it's listed in all this information, is showing you that visually this is what's going on as well</u> .	- Description of previous statement - She would like to receive provenance information as a combination of tabular format with visual effects.	PRESENTATION Tabular format Visual effects	10	7



Methodology 1 – Important sub-themes

IMPORTANT FACTORS PRESENTED THROUGH PROVENANCE	Sub-themes	Representative quotation	Why this quote fits the theme?	Par/Ref (Total)
	Quality Considerations	"Some of them are obvious, like completeness, validity, usage, currency...Listen to the other ones, I would not really have known about them or considered." (Participant 11, p. 7)	It demonstrates the basic understanding of end users to quality issues.	8/20
	Source	"For me it's all about that, that authorised source. (Participant 5, p. 6)	Source (authorised or not) should be known.	9/24
	Date related	"...the data is very messy, and nobody likes to work without a date" (Participant 10, p. 4)	It demonstrates the importance of knowing the date.	11/59
	General Considerations	"Eh, just give you a bit more information about the data. It's nice to have, is not essential though." (Participant 2, p. 8)	General information can improve the understanding.	7/11
	Identification	"... if you say try to add them like date when it's collected then by who? Lie by who, but not by person, by each organisation..." (Participant 8, p. 7)	It appears that knowing the organisation of where the data collected is important.	10/39
	Processing Considerations	"And it is important to know, to know why as well. Um, I suppose that kind of links in with...what they done to it and why have they done it?" (Participant 5, p.6)	The need to know more about the process of the information and the intended use of it, exposes.	6/15
	Security	"Um, and then, uh, the, uh, the classification. So, who can they be released to? Classification, uh, slash releasability... You difficult to make a decision with it if you can't push it out there to the commander for example, if he did not have a security clearance to see it." (Participant 3, p.8)	It also shows a striking example of the importance of security classification.	6/7
	Spatial related	"It's good having a positional accuracy but if I don't know the reference system that was used to create it then my understanding of the accuracy of position may be different to the creators." (Participant 1, p. 7)	A mixture of spatial elements may be needed in that case.	7/25

Methodology 1 - Results



Methodology 1 - Results

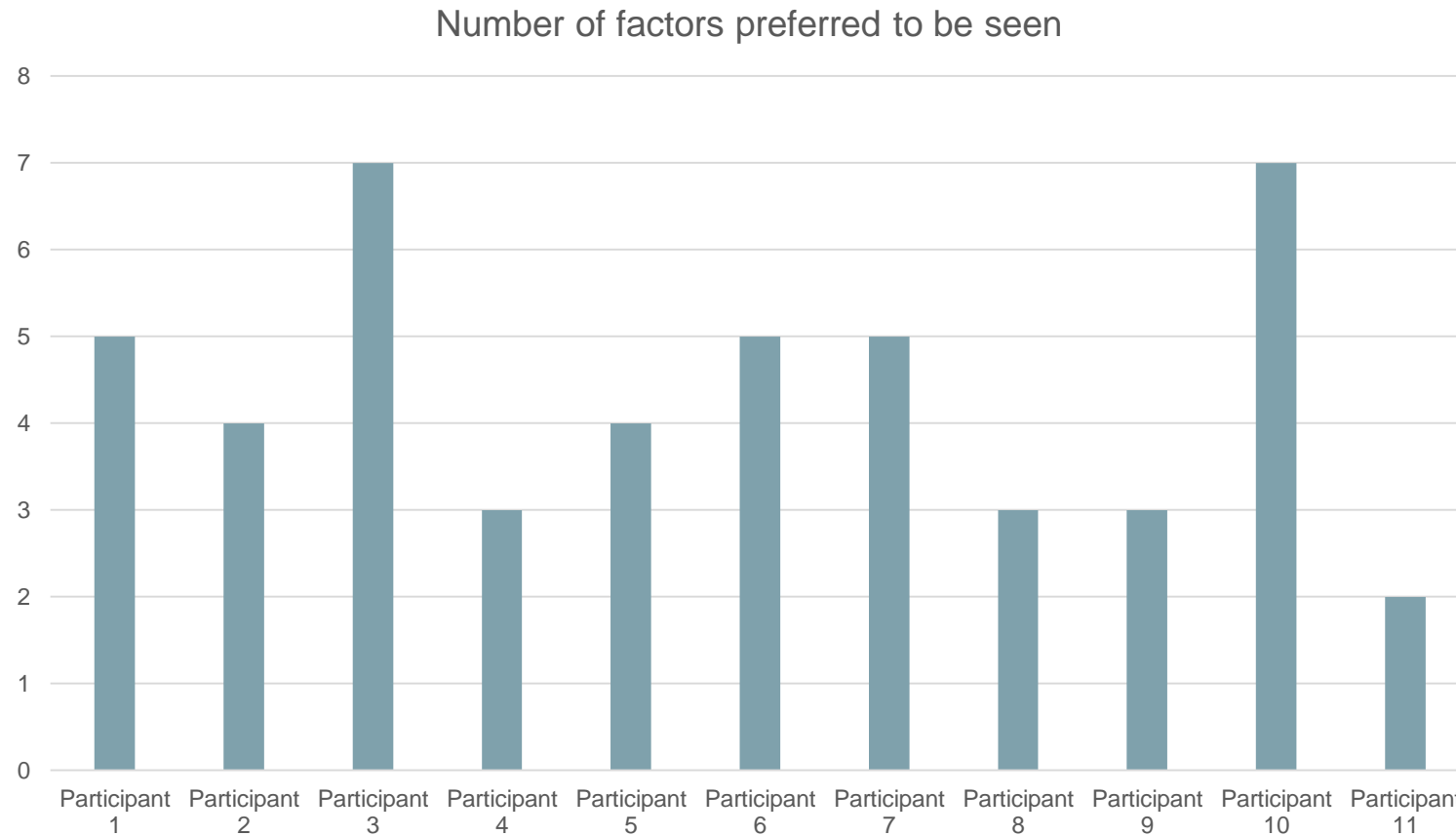
Participant	Ranking a date related factor as the most important to be seen.
Participant 2	"Yeah, the date captured 1..."
Participant 3	"...you could probably shift them around, but the top few definitely currency, completeness and the data collection source and resolution is, you know, is definitely important."
Participant 4	"You'd always wish for the most up to date...piece of data."
Participant 5	"So, when, you know, currency is everything these days... Dates, it's all about dates and who."
Participant 8	"...if I'm thinking about kind of metadata, we try always to attach to our data sets as the source and the date of the, of the source, like the data that is created."
Participant 9	"Anything with a date on it is very important."
Participant 10	"...the biggest, biggest thing is just from my colleagues' point of view is understanding the date, uh, first glance."
Participant 11	"The key elements, you know, it's the, you know: How recent the data is and uh, you know, the source of that data."

Methodology 1 - Results

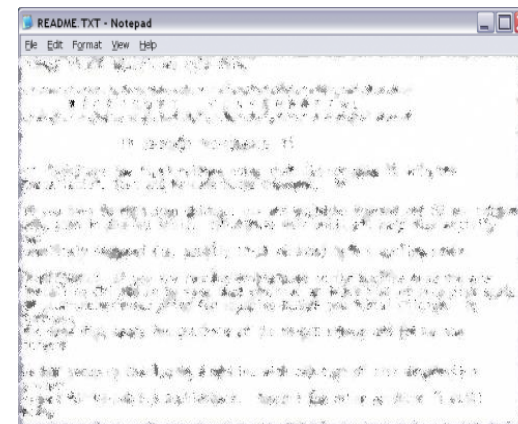
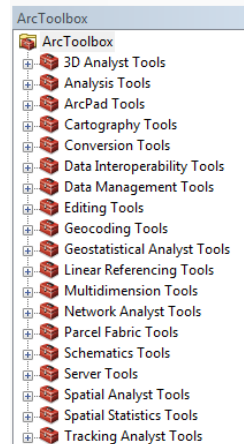
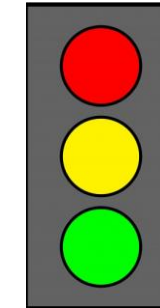
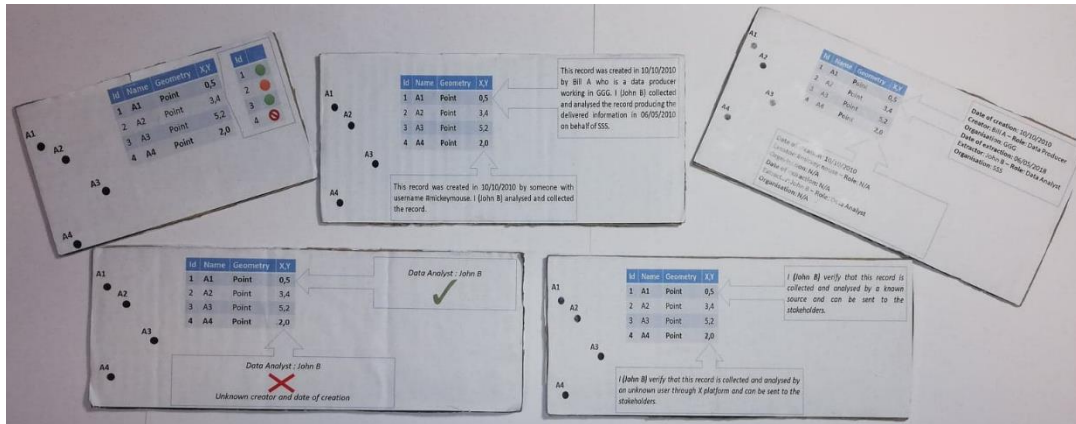
- “...to be honest, I wouldn’t, yeah, I probably wouldn’t necessarily be, eh, fully aware of that. I’m not, **I’m not an analyst**, so, so, the, yeah the accuracy of that data in those sort of parameters will be not relatively caught previously.” (Participant 11)
- “Anything with a **date** on it is very important.” (Participant 9)
- “But the most fundamental things, again, from my assessment, I would like to be able to see those **straightway** that kind of staff. To give me a **quick understanding** and not just me but also the customers...” (Participant 1)

FACTORS/PARTICIPANTS	1	2	3	4	5	6	7a	7b	8	9	10	11	TOTAL	MODE
DATA QUALITY														
Completeness			2										1	2
DATA SOURCE														
Data Source	4		3				1	4		3		2	6	3, 4
DATE RELATED														
Date Produced	2	1		1	1	2	3	1	1		1		9	1
Date of analysis					3		3	1					3	3
Date Valid until	5						3	1					3	N/A
Currency			1	2			3	1		1	3	1	7	1
GENERAL														
General description											6		1	N/A
Type of data						3							1	N/A
Intended Usage			5			4							2	N/A
Metadata (General)											2		1	N/A
IDENTIFICATION														
Organization (name)	4				2	1			3	2	7		y	2
Contact							2	5					2	N/A
PROCESSING														
Method used					4		5	2					3	N/A
SECURITY														
Security Classification	1		6			5					5		4	5
SPATIAL RELATED														
Positional Accuracy	3			3							4		3	3
Scale/Resolution		3	4										2	N/A
Spatial Reference			7										1	N/A
Location		4											1	N/A
EMERGENT														
Accuracy (General)		2					4	3	2				4	2

Methodology 1 - Results



Methodology 1 - Results

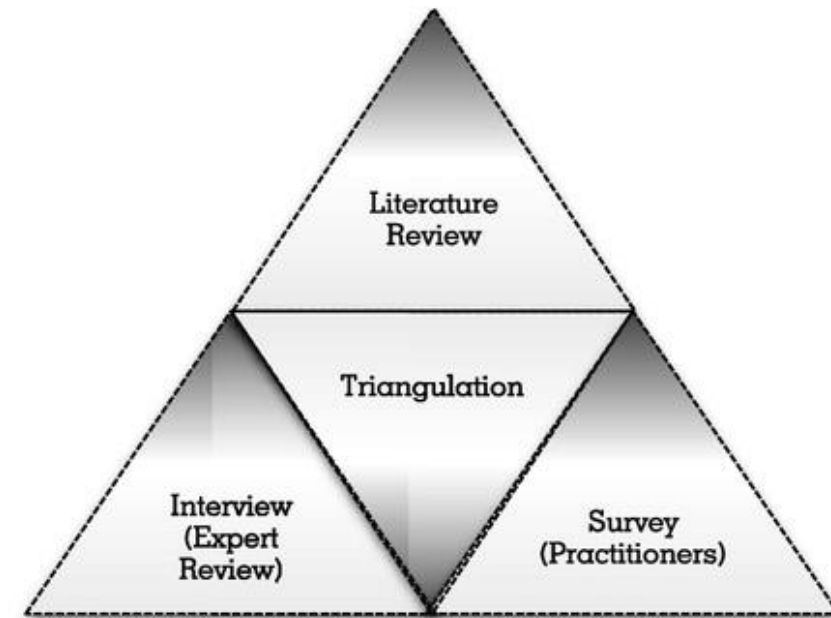


OID	ForestID	FireName	Year	Month	Day	Date	DispatchFr	ReturnTo	Area	
0	WA-OWF	797	2015	9	15	9/15/2015	NCSB	NCSB	Northwest	F
1	WA-OWF	683	2015	8	15	8/15/2015	NCSB	NCSB	Northwest	F
2	WA-OWF	684	2015	8	15	8/15/2015	NCSB	NCSB	Northwest	F
3	WA-OWF	639	2015	8	13	8/13/2015	NCSB	NCSB	Northwest	F
4	WA-OWF	505	2015	7	20	7/20/2015	NCSB	NCSB	Northwest	F
5	WA-OWF	517	2015	7	20	7/20/2015	NCSB	NCSB	Northwest	F
6	WA-OWF	510	2015	7	20	7/20/2015	NCSB	NCSB	Northwest	F
7	WA-OWF	502	2015	7	20	7/20/2015	NCSB	NCSB	Northwest	F
8	WA-OWF	500	2015	7	20	7/20/2015	NCSB	NCSB	Northwest	F
9	WA-OWF	Ugularo/Johansen	2015	7	20	7/20/2015	NCSB	NCSB	Northwest	F
10	WA-OWF	495	2015	7	20	7/20/2015	NCSB	NCSB	Northwest	F
11	WA-OWF	Junction Mtn	2015	7	11	7/11/2015	NCSB	NCSB	Northwest	F
12	WA-OWF	409 Libby Creek	2015	7	10	7/10/2015	NCSB	NCSB	Northwest	F
13	WA-OWF	410 OWF	2015	7	10	7/10/2015	NCSB	NCSB	Northwest	F
14	WA-OWF	Splawn Crk 402	2015	7	9	7/9/2015	NCSB	NCSB	Northwest	F
15	WA-OWF	War Creek 396	2015	7	8	7/8/2015	NCSB	NCSB	Northwest	F

These photos are licensed under [CC BY-SA-NC](https://creativecommons.org/licenses/by-sa/4.0/)

Future Plans – Validation (Internal and External)

- **Data collection method** (Methodological Triangulation)
 - Literature review
 - Semi-structured Interviews
 - Online Questionnaires



Alassafi et al. 2017

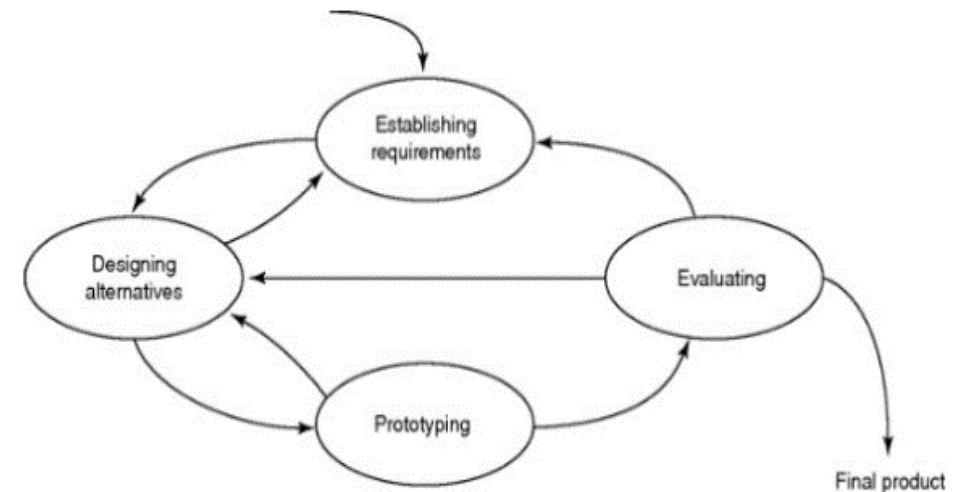
Future Plans – Methodology 2

- **Prototyping**

- Low fidelity: simple, cheap and quick
- High fidelity: higher level of functionality, approach final product

- **Evaluation**

- Collect information about decision-makers' interaction
- Measure their performance
- The evaluation method depends on the fidelity level of prototyping



Preece et al. 2015

Presenting and evaluating provenance information (Usability and trust)

User Testing	Inspection (replacing user feedback)	Field
Think aloud	Expert Judgment	Behavioural observations
Metrics	Guidelines and checklists	Collages or artefacts
Post-use usability questionnaires	Heuristic Evaluation	Log analysis
Engagement	Walkthroughs	Experience sampling method (ESM)
Aesthetics		Living laboratory
Interviews/Focus groups		
Emocards		
Personal meaning maps		
Facial expressions		
Physiological reactions		

Table produced by Macdonald and Atwood (2013)

Future Plans – Needs

- ! Do not hesitate to contact me you are interested to take part in the next stages of the research (online survey, usability testing) !

nikolaos.papapesios.16@ucl.ac.uk

Related work

- Papapesios, N., Ellul, C., Shakir, A. and Hart, G., 2019. Exploring the use of crowdsourced geographic information in defence: challenges and opportunities. *Journal of Geographical Systems*, 21(1), pp.133-160.
- Noskov, A., Grinberger, A.Y., Papapesios, N., Rousell, A., Troilo, R. and Zipf, A., 2019. Modelling and Assessing Spatial Big Data: Use Cases of the OpenStreetMap Full-History Dump. In *Spatial Planning in the Big Data Revolution* (pp. 16-44). IGI Global.

Thank you

Nikos Papapesios

PhD Researcher

University College London

Civil, Environmental & Geomatic Engineering

First Floor - Chadwick Building

Gower Street

London WC1E 6BT

mob. +44 7901837318

email. nikolaos.papapesios.16@ucl.ac.uk