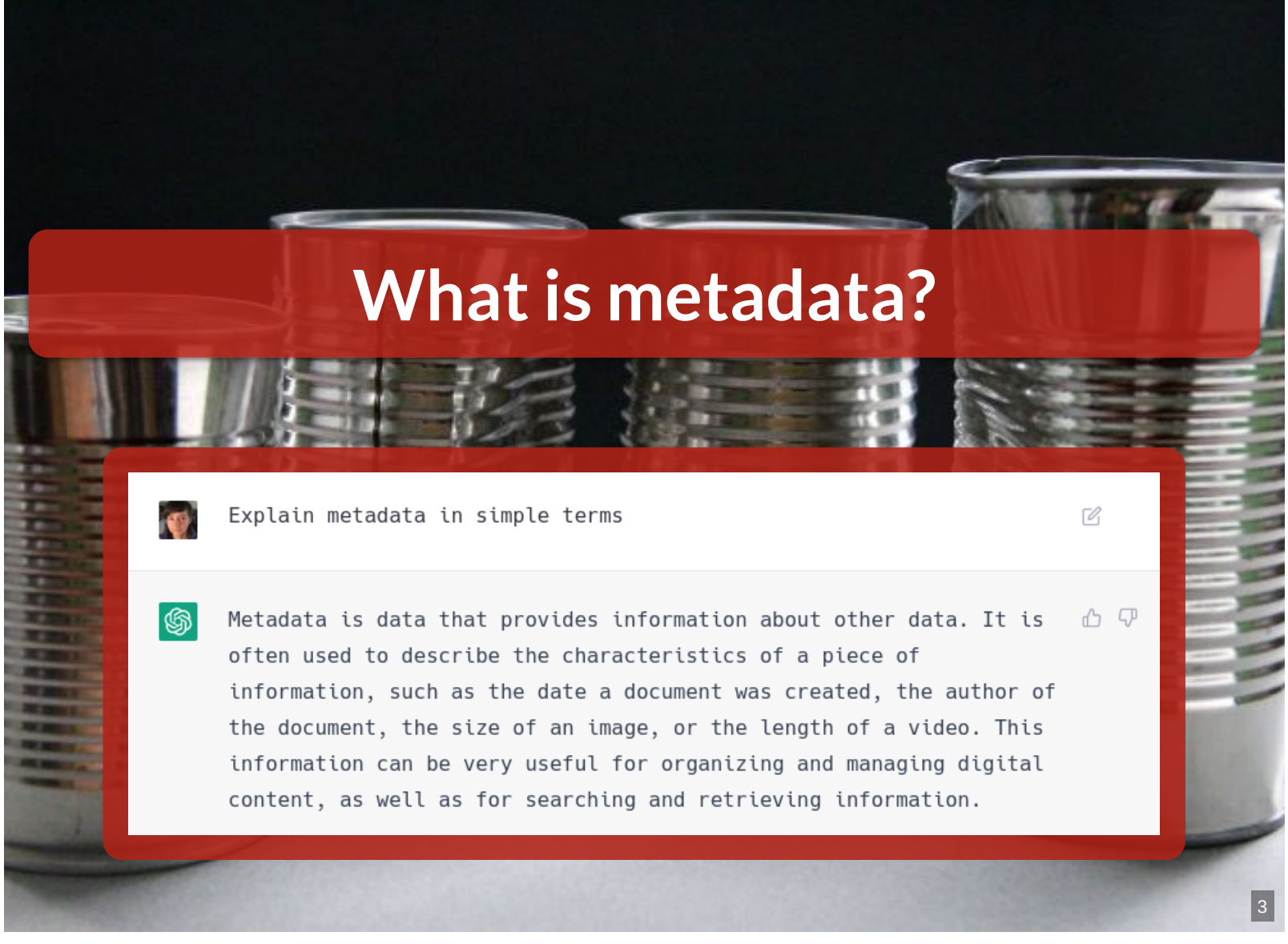
A Beginner's Guide to Metadata

Jo Cook | Astun Technology


schema.org
dcat
csw
xsd
rdf
schematron
xml
data quality framework
knowledge graph
bsi
iso19139
iso19119
triple
datacube
xquery
json
iso19110
iso19157
gemini
path
feature catalog
linked data
dcat-ap
dublin core
Hello! 👋
Welcome to a talk by the Data Discoverability team at Astun Technology

Speaker notes

We work with GeoNetwork and other technologies to make it as easy as possible for people to create good-quality metadata without being overwhelmed by what is without a doubt a complex and technical subject. Over the next 45 minutes I'm going to outline what metadata is, why it's important, highlight some of the jargon you might come across, and set out some easy steps you can undertake to get started


What is metadata?
Explain metadata in simple terms
Metadata is data that provides information about other data. It is often used to describe the characteristics of a piece of information, such as the date a document was created, the author of the document, the size of an image, or the length of a video. This information can be very useful for organizing and managing digital content, as well as for searching and retrieving information.

Speaker notes

Metadata is usually described as data about data, but that doesn't really tell you much. I asked my good friend (on the internet) ChatGPT for advice.

Why is it important?

summarise why metadata is important

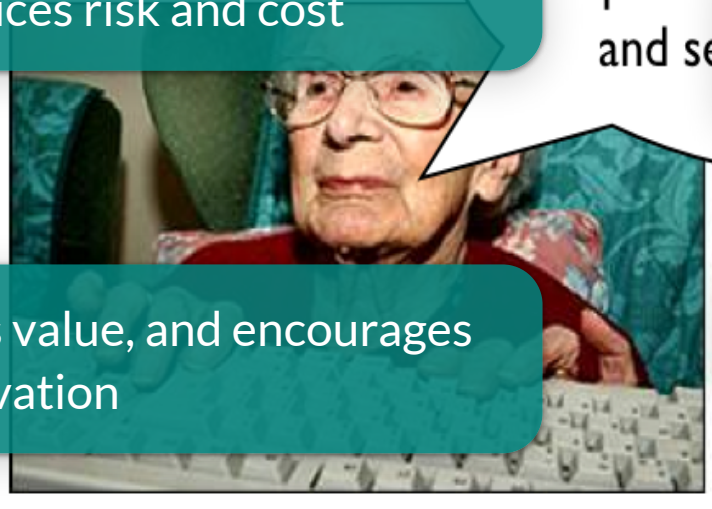
Metadata is important because it provides information about the content and context of a piece of data, allowing users to understand and interpret it correctly. It also enables effective management and organisation of data, facilitating discovery and retrieval. In addition, metadata can support security and privacy by providing information on the access and use of data. Overall, metadata plays a crucial role in the effective use and preservation of data.

Speaker notes

So chatGPT knows that metadata is important for making data management more effective, secure, and efficient

But it also...

Reduces risk and cost



and helps others understand the data

Adds value, and encourages innovation

👉 Serious example

Speaker notes

COVID- we all got pretty good at deciding whether case data was useful or not. For this sort of dataset you need to know its getting daily updates. Metadata reduces the amount of time people need to search for datasets, and reduces the likelihood of their choosing the wrong dataset for the task at hand. A good catalogue will also be optimised for discovery, for humans and search engines, and will include user statistics and feedback options, helping data providers know which datasets are being used, and providing an easy way for people to report problems.


However...
Why is metadata hard?
Metadata is often difficult because it can be complex and hard to understand. It can also be difficult to organize and manage, especially when dealing with large amounts of data. Additionally, metadata often has many different standards and formats, making it challenging to ensure that it is consistently collected and organized.

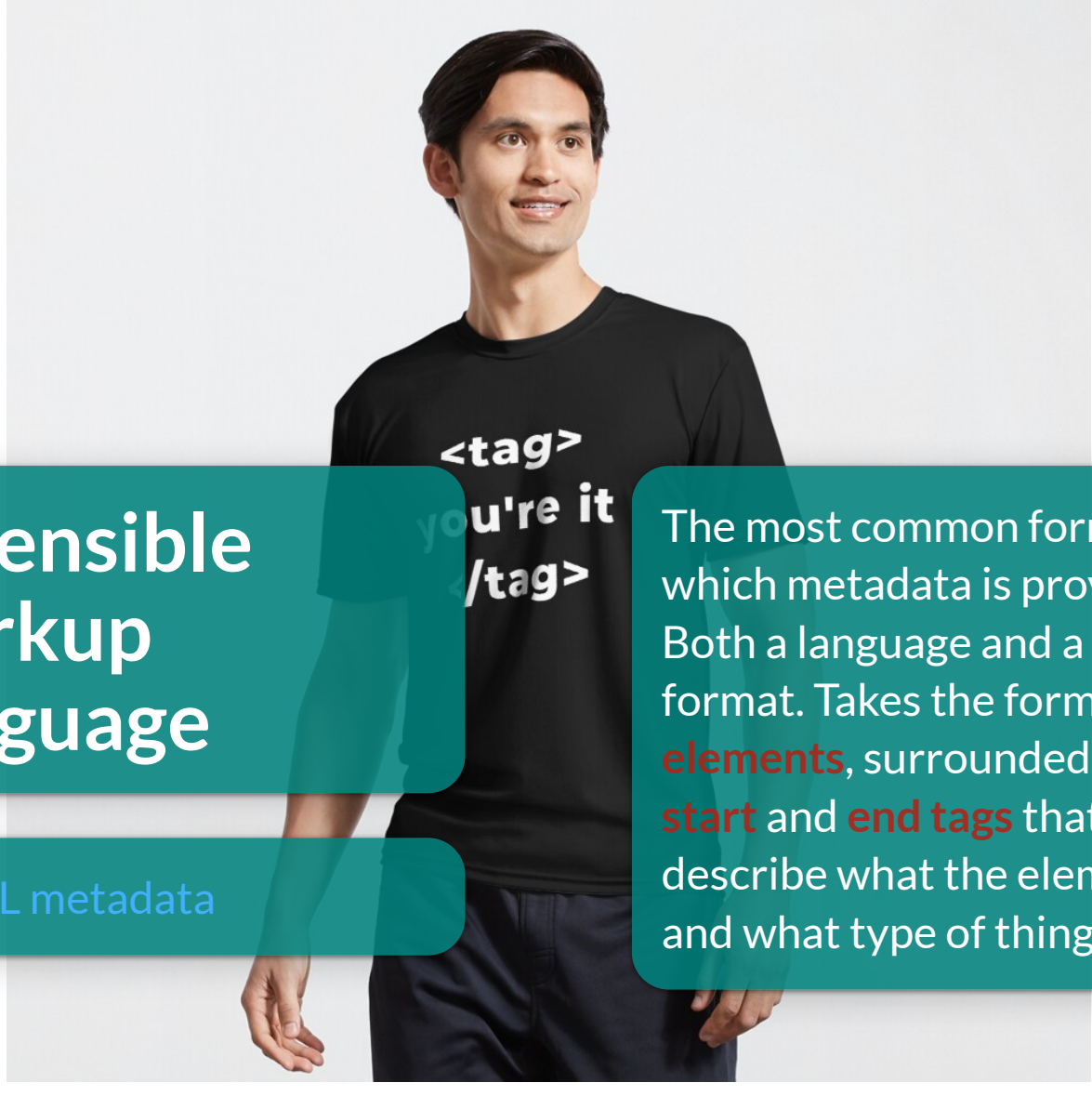
Speaker notes

As ChatGPT said, metadata is complex, difficult to organise, with many different standards, formats and terminology. Hopefully we can make some of this easier to understand in the rest of this talk!


Hello
Some jargon
Can I talk to you
about metadata?

Speaker notes

In this section, we'll cover some of the most common acronyms and terms that you might hear when working with metadata, hopefully with a simple explanation



eXtensible Markup Language

👉 XML metadata

The most common format in which metadata is provided. Both a language and a file format. Takes the form of **elements**, surrounded by **start** and **end tags** that describe what the element is and what type of thing it is

Speaker notes

Metadata is often supplied in xml format because of the flexibility and extensibility of the format. It's both machine and human readable (with practice). A good metadata catalogue will shield you from most of the complexity

XML rules and validity

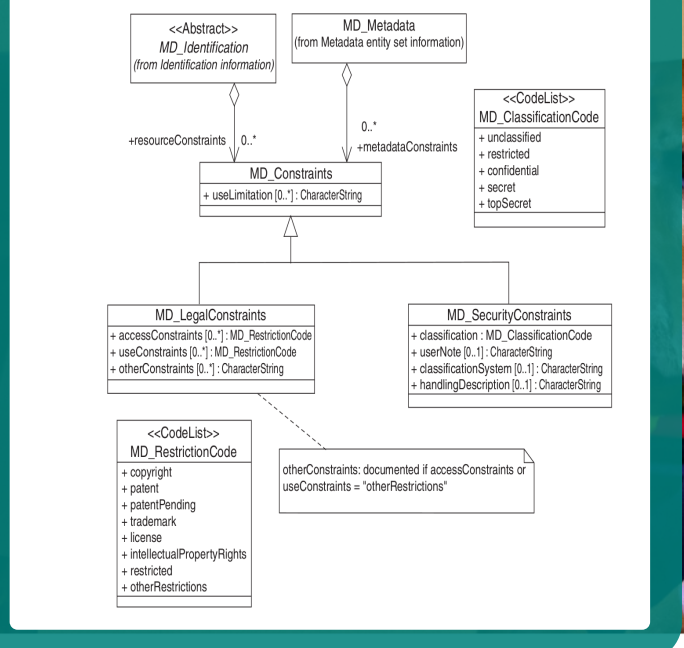
Rules on what should be in a metadata record are defined by **schemas**. A record can re-use elements from many different schemas using **namespaces**. A valid record must be both **well-formed** and **schema-valid**

👉 That record again

Speaker notes

Schemas are basically frameworks for what should be included when creating metadata. To avoid re-inventing the wheel, schemas will use elements from other schemas to describe certain concepts. The different schemas are referred to in the xml using the namespaces. Valid metadata must be both xml-valid (so no dangling tags, for instance) and schema-valid. There are a number of other concepts, such as XSL(T) for transforming XML, XPath for referring to a particular bit of XML in a document, and XQuery for doing SQL-like queries. These are used extensively within metadata catalogues such as GeoNetwork to control the display of records on the page, and to allow updates to records.

Schemas, Standards and Profiles



Schema: specifies the structure of the metadata, the elements, and the controlled vocabularies used. **Standard:** a schema that has been developed and maintained by a standards organisation. **Profile:** an extension to a standard, or an implementation of it in software, or both

Speaker notes

You'll hear all of these used interchangeably when people talk about metadata. Generally speaking, a profile can only be more strict than the standard it extends. This might be by making an element mandatory when the core standard says it's optional, or by adding a controlled vocabulary. Elements from other standards can be added as long as they don't conflict.

ISO Standards and Gemini

"a picture of a dog looking sleepy while staring at a computer screen showing code, in the style

Jo × DALL·E

The core standard for Geospatial metadata is ISO19115. ISO19139 is the implementation of ISO19115 in XML. GEMINI is an extension to ISO19139 for UK geospatial metadata

👉 Gemini Online

Speaker notes

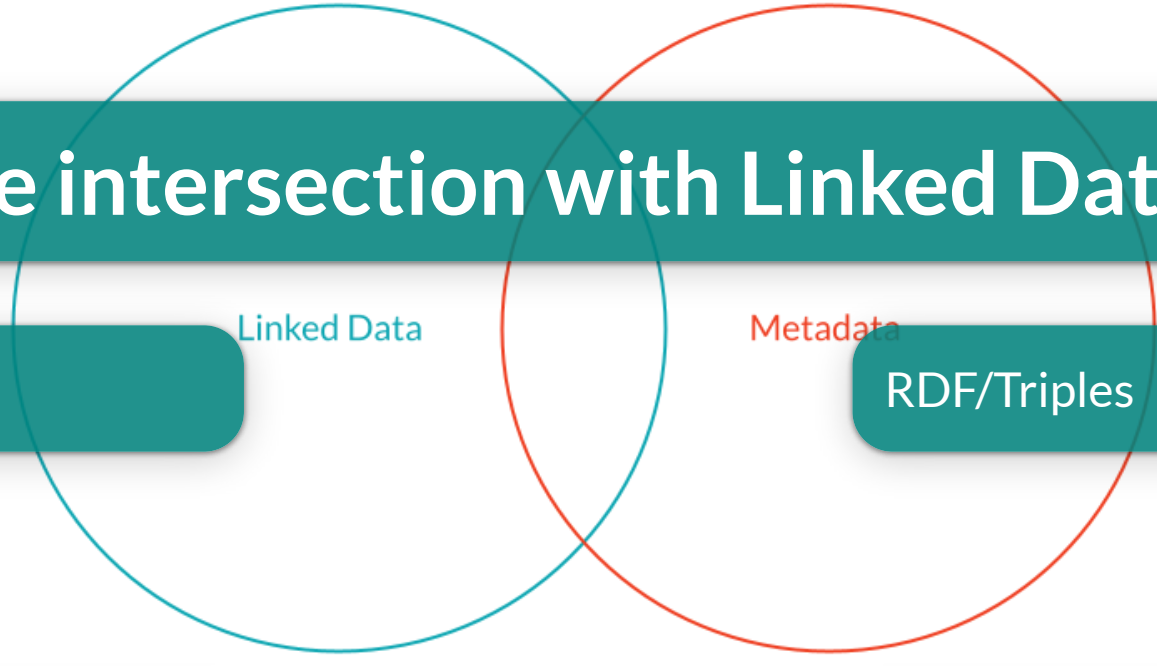
There are a couple of official ISO standards that you'll commonly encounter. Generally you can just focus on UK Gemini, where the implementation work has been done for you. It adds in some UK-specific terms and code lists, reduces some choices (such as language), and adds some elements from other standards such as ISO19119 for services


What about INSPIRE?
#1 Identify key environmental datasets
#2 Provide metadata in Gemini format
#3 Make data available via web services
#4 Publish to data.gov.uk
12

Speaker notes

INSPIRE was an EU regulation for the sharing of environmental spatial data, originally implemented in 2007, and brought over into UK law after Brexit. So yes, you still have to do it, but it's easy as a valid Gemini record is also INSPIRE-compliant, you simply have to include a keyword from the INSPIRE spatial data themes

The intersection with Linked Data



URIs

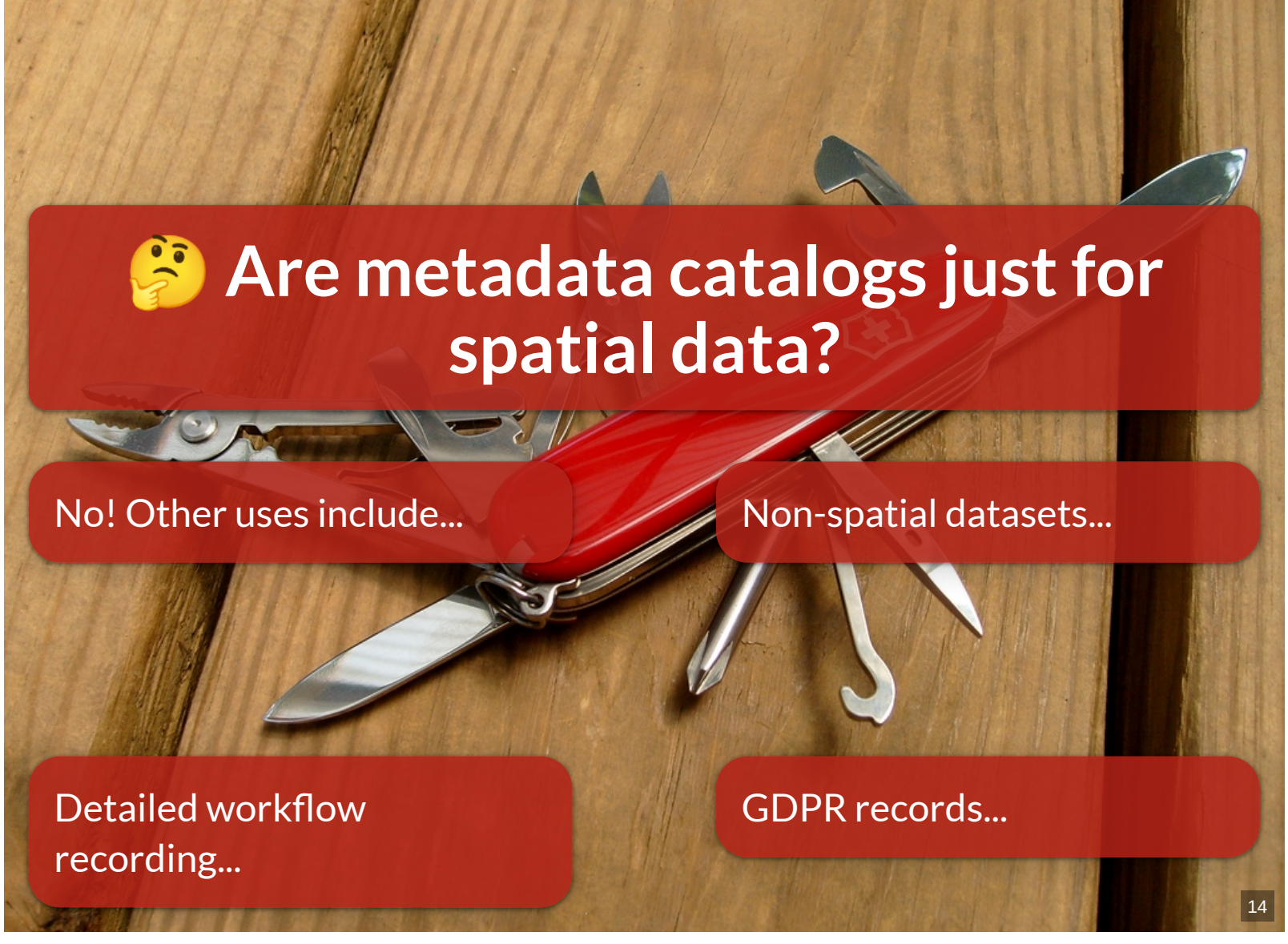
RDF/Triples

Dublin Core/DCAT

Ontologies

Speaker notes

If metadata is data about data, linked data is data structured so that it can be easily connected and integrated with other data. Linked data uses standard web technologies such as URIs (universal resource locators, of which a web address or URL is a subset) to identify and make data available in a machine-readable way. Other terms that you might come across at the intersection of metadata and linked data are: Dublin Core is a basic, domain-agnostic standard for describing any sort of resource. DCAT and (Geo)DCAT-AP are profiles of Dublin Core, widely used in catalogues such as CKAN. RDF is a data model that comprises "triple statements", comprising a subject, a predicate or relationship, and an object, and are widely used for linked data and semantic web implementations. The Dublin in Dublin Core relates to Dublin, Ohio rather than Dublin, Ireland, and that is a triple statement where the subject is "Dublin Core", the predicate is "relates to", and the object is "Dublin Ohio". An ontology is a fancy word for a set of categories or concepts within a domain. A good metadata catalogue will allow you to publish and ingest metadata in both ISO19139 and DCAT formats.


🤔 Are metadata catalogs just for spatial data?
No! Other uses include...
Non-spatial datasets...
Detailed workflow recording...
GDPR records...

Speaker notes

The core ISO19139 standard allows the use of elements from many other standards, so you can extend records to store additional information to meet your needs, in a structured, standards-compliant, interoperable way.

What is Q-FAIR all about?

Quality

- is just a hyphen

Findable

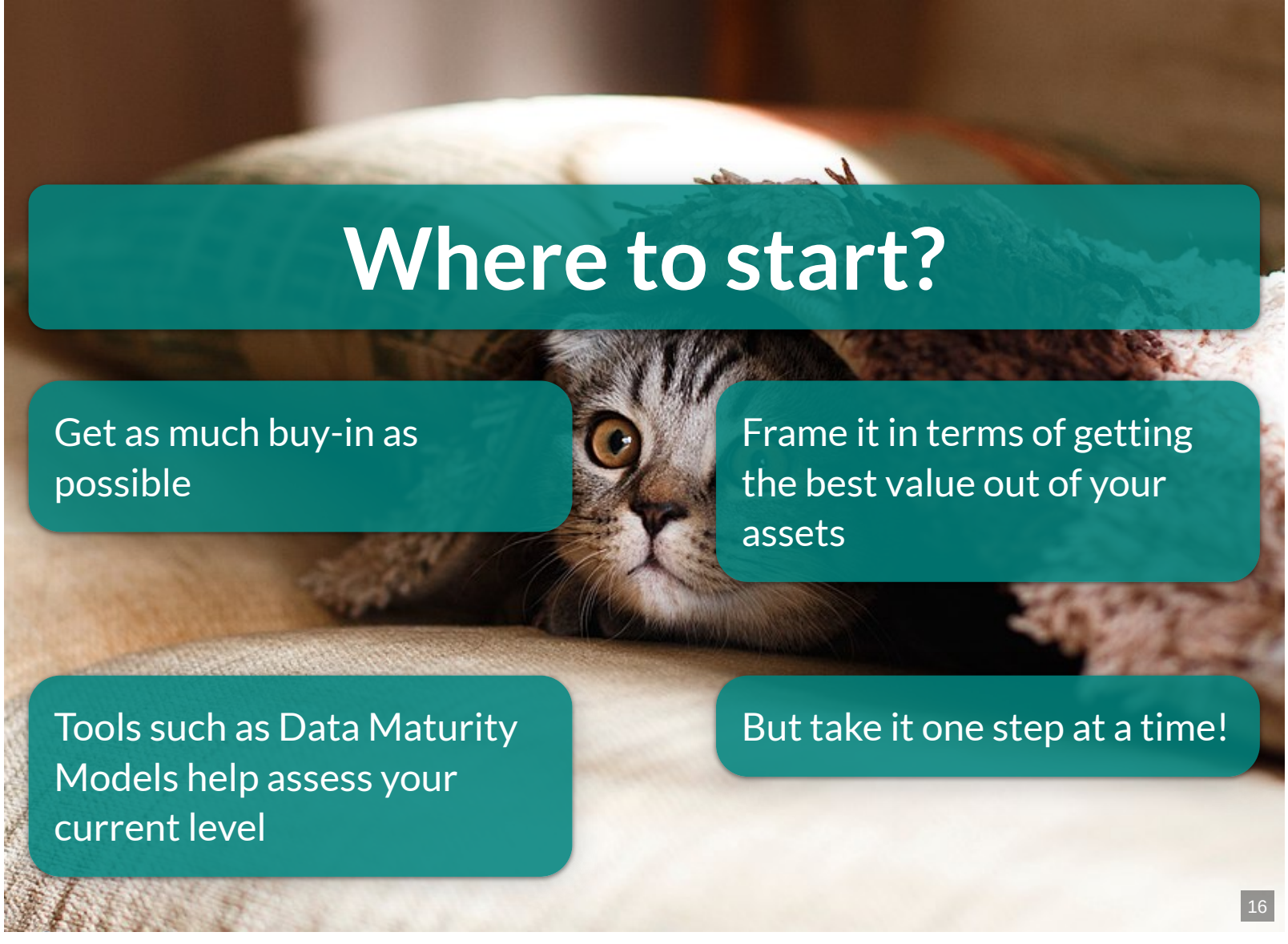
Accessible

Interoperable

Reusable

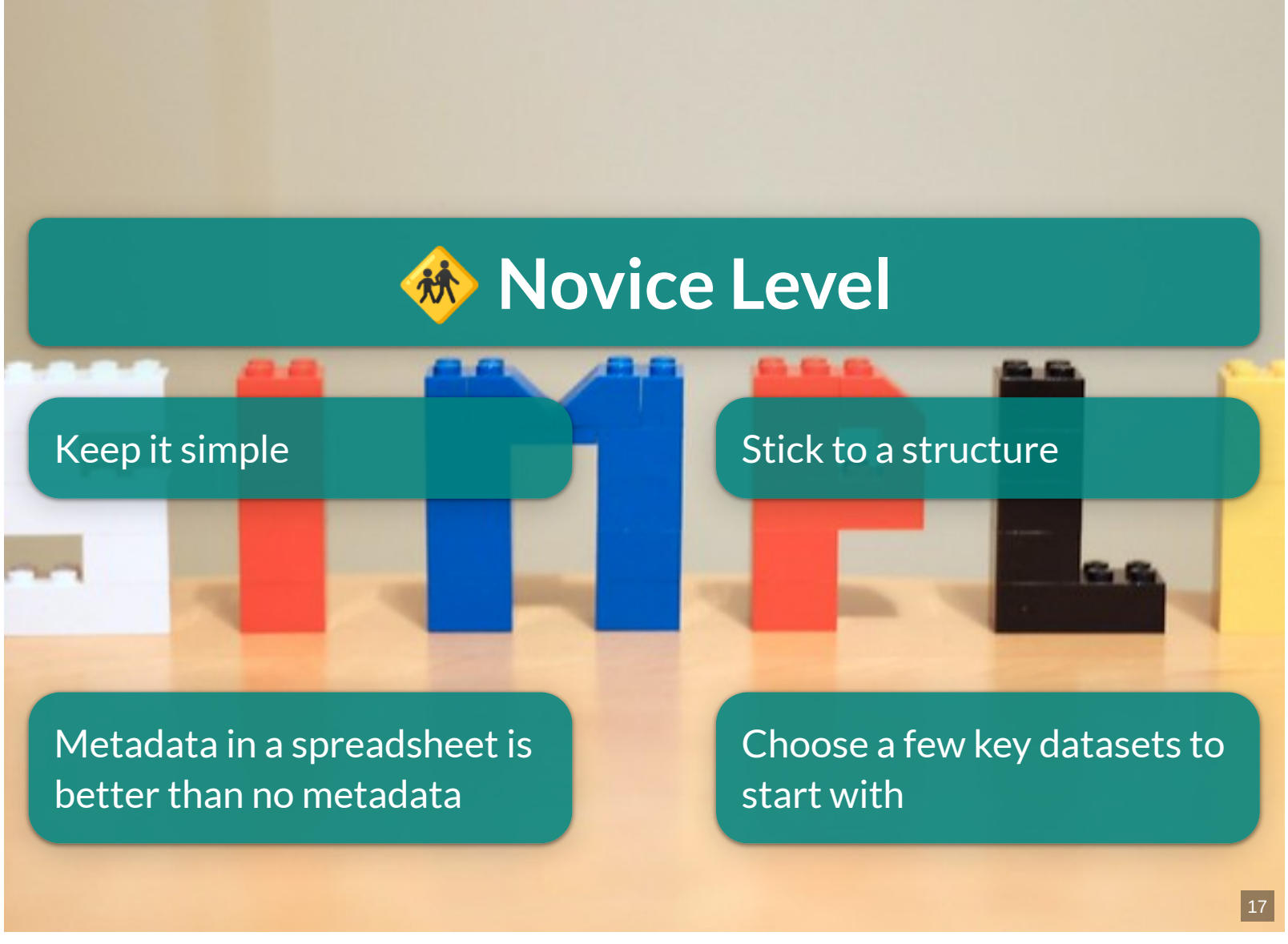
Speaker notes

Q-FAIR is all about making your data high-quality, findable, accessible, interoperable and reusable. This is best (or most easily) achieved if your metadata is also Q-FAIR, and also best achieved using open standards and formats where possible.


Where to start?
Get as much buy-in as possible
Frame it in terms of getting the best value out of your assets
Tools such as Data Maturity Models help assess your current level
But take it one step at a time!

Speaker notes

After the jargon, here are some simple tips for how to get started. Go back to the benefits of data sharing to get as much buy-in as possible, and use tools such as Data Maturity Models to identify where you should focus time and energy. It can be done a bit at a time though, and here are some simple ways to get started


Novice Level
Keep it simple
Stick to a structure
Metadata in a spreadsheet is better than no metadata
Choose a few key datasets to start with

Speaker notes

Recording a small number of key metadata elements for your datasets in a simple spreadsheet is a good way to start. It's always possible to script the insertion of them into a catalogue if you choose to

💪 Next Level

Don't reinvent the wheel

Adopt open solutions that aid sharing and discovery

Always be Q-FAIR

Don't let perfect be the enemy of good

Speaker notes

There are many metadata solutions out there, you don't need to craft your own. To make your life, and those of people using your data, choose a solution that adopts open standards and promotes a Q-FAIR approach. Don't worry if your metadata or data are not perfect though- getting something out there is much better than nothing!


HELP
SUPPORT
Whatever stage you're at...
We can help! 👉 astuntechnology.com/dd


Any questions?


🙏 Thanks for attending!
We'll be sharing the recording with you all shortly