Hi, I'm Jack. I co-founded Open Climate Fix and I spend half my time consulting for National Grid ESO. Just to be clear: I'm writing this document with my "Open Climate Fix" hat on so this document in no way attempts to speak for ESO! But I spend a lot of my time thinking about how to help the ESO, so a lot of this document is seen through the lens of helping operate the electricity grid because that's what I have most experience with!

I've deliberately written this doc in concrete language, where I make specific proposals. That's not because I want to dictate how this will work! I fully expect all of the ideas in this document to change radically as we discuss them! But I think it's useful to write in concrete terms so people can point at specific statements and say "that won't work because...". So please comment liberally on this doc!

Energy data should enable rich digital twins

My favourite film is the anti-war 1983 classic, WarGames. One of the main locations in WarGames is NORAD: the NORth American aerospace Defense command. It's a huge room filled with screens giving a real-time, super-detailed view of America's defense systems. NORAD is all about making America's defense data visible:

Let's do the same energy: Let's share enough energy data to enable innovators to build gorgeous, interactive, real-time "digital twins" of the entire energy system which display multiple layers of energy data.

This "digital twin" wouldn't be part of the EDVP itself, but the energy metadata defined by the EDVP should enable other innovators to build a map UI. Crucially, it's important to consider the 'digital twin' use-case now so we ensure the EDVP's metadata enables feature-rich use-cases like digital twins. At the very least, the metadata should capture spatial information, and should describe datasets in sufficient detail to enable the automatic ingestion of diverse datasets into a 'digital twin', and should describe relationships between assets and datasets.

These digital twins would go far beyond a static map of assets: These digital twins would also show the current state of each asset and their relationships to other assets. Click on a generator and the UI will highlight the wiring between the generator and the local substation, and show real-time power flowing from the power station. Select specific 'layers' of data (e.g. only show the solar systems or the transmission lines). Show network congestion just like how Google Maps shows road traffic congestion. Animate energy flows across time and space like this animation of solar PV generation across the UK. Use physical modelling to infer the state of assets for which we have no direct data. Enable users to run forecasts / simulations for many different scenarios of the future. Go back in time to study the state of the energy system in the past and to run "what if" scenarios.

The EDVP Beta (or the EDVP Alpha 'developed' MVP)

There could be 4 main interfaces to the EDVP:

- Textual search
    - Run rich semantic searches like "Show me all the time-series datasets which record solar PV power output at hourly resolution or better; and cover the south-east of England from 2015 to 2020. The data must be from individual PV systems instead of being aggregated; and must be released under a permissive open license like CC-BY-4.0."
- Browse a hierarchy of categories (e.g. energy → electricity → generation → solar PV)
- Map search
    - Click on an asset to see all the datasets associated with that asset
    - Click-and-drag to see all datasets for a geographical region
    - See the geographical distribution of datasets that satisfy a textual search (e.g. show me the spatial coverage of each solar PV dataset with data for 2020)
- API

Under the hood, the energy data visibility system will construct a knowledge graph (described in this blog post) which represents the entire energy system: the assets, the datasets, and their relationships. Once this knowledge graph is constructed, it's relatively easy to expose that data however the user wants.

The source data for the knowledge graph will be decentralised, and other people will be free to build their own knowledge graphs. The source data will consist of small, machine-readable files published on the web servers of the country's energy organisations. These files will link to each other (just as web pages link to each other) to represent relationships. These files will describe energy assets, energy datasets, and the relationships between all these entities. These files will constitute a "web of energy data".

These metadata files will be written using a standard format and a standard vocabulary. The vast majority of users won't create metadata files from scratch in a text editor. Instead, the metadata will be created a little like how HTML is created today: Most HTML on the web was created by people who have no idea how to write HTML from scratch. Instead they use tools (mostly on the web!) to create HTML (e.g. to write a blog, use Medium.com).

The technical standards for the metadata and datasets will be evolved rapidly by the community out in the open, just as open source software is designed by distributed teams in the open. Guardrails will be set by Ofgem but, within those guardrails, the community is free to update the technical standards as they see fit using simple online discussion tools.

Why do we need better data?

I spend half my time at National Grid Electricity System Operator, so I'll focus on operating the grid, but there are a huge number of other use-cases for sharing energy data.

Operating the electricity grid today is a little like trying to remotely manage a large team over a really glitchy, laggy, low-resolution video conferencing system, where half the team haven't even bothered to turn up.

As the manager, you also have an enormous responsibility: the cost of failure is truly catastrophic (the grid going down). So, to make up for the rubbish communications links, the grid operator has no choice but to run the system with tonnes of contingency (mostly 'spinning reserve'). In general, when your ability to observe and control a fast-moving, complex system is impaired, you can't run that system anywhere near its optimal performance. You have to run the system with loads of slack.

On the electricity system, that 'slack' is very costly: It cost over a billion pounds to balance the UK grid in 2020 (that cost is paid by bill payers). And there's a huge carbon penalty too, because a lot of the spinning reserve comes in the form of fossil-fuel powered generators.

We could substantially reduce the costs of balancing the grid and substantially reduce carbon emissions 'just' by improving information flows. Better situational awareness means we can run the system closer to its optimal performance (with less slack). In fact, I'd argue that if we're to achieve net-zero, we *have no choice but to* get a lot better at sharing data. A net-zero grid will be far more complex than today's grid, and so will *absolutely require* far better information flows. Net-zero isn't possible unless we get a lot better at sharing data!

I can hear some people shouting "but the electricity system operator already has all this data!" So let me give some concrete examples of the ways in which the data feeds to the Electricity System Operator are broken today. And, to be clear, this isn't the ESO's fault: The ESO have to make do with whatever data they're given by the wider system. And these problems are present on grids around the world. In fact, the British grid is doing better than many!:

- The control room doesn't receive data recording the nation's true electricity demand (the total power demand of the nation's kettles, lights, computers, offices, factories, etc.). That's really weird, given that the control room's *main job* is to balance electricity supply & demand. Yet they don't receive good demand data. Yes, the control room gets real-time telemetry from grid supply points (the boundary between the transmission and distribution systems), but there's increasingly large amounts of 'embedded' generation: generation which, from the ESO's perspective, is 'behind'

the grid supply points. Meter readings from each grid supply point represents total demand minus an unknown amount of embedded generation.

- The control room doesn't get live data from a large proportion of the nation's power generators. Again, remember that the control room's *main job* is balancing supply and demand! It's like we're asking them to do their job blindfolded! The control room gets absolutely no realtime data for most embedded generation (with the exception of PV, where the control room gets a real-time estimate for the nation's 1 million PV systems. This estimate is inferred by the clever folks at Sheffield Solar from a real-time feed from about 1,000 PV systems. Those 1,000 PV systems represent 0.1 % of the nation's PV fleet). One of my jobs whilst consulting for ESO is to pull in Electralink's half-hourly MPAN data from embedded generation. But the Electralink data isn't realtime: it's at least 24-hours old. And the Electralink data doesn't capture data from meters which export less than 30 kW to the grid, so it doesn't capture domestic microgeneration. Oh, and the MPAN admin database doesn't record fuel type, so we struggle to figure out what type of microgeneration is behind each MPAN meter. The Embedded Capacity Register definitely helps, but the ECR doesn't capture DERs under 1 MW. Embedded generation breaks down (very roughly) as:
    - 13 GW of solar PV (with another 10 GW in the planning pipeline)
    - 7 GW of wind generation
    - 5 GW of thermal generation (CHP, gas turbines, diesel, etc.). This thermal generation is particularly worrying for the control room because a lot of it is price-sensitive (and the UK has a single price electricity). So, even though each individual unit might be small, large numbers of these generators turn on and off in unison in response to swings in the electricity price. So the control room might see multiple GW of embedded thermal generation magically disappear with absolutely no warning. This is bad.
    - 1.1 GW of batteries (with another 16 GW in the planning pipeline)

A system operator's main job is to ensure, at every moment, that electricity supply = electricity demand. Yet they don't receive reliable data for either side of this equation! So they have to run the system with loads of slack, and use the system frequency as the authoritative guide to the balance of supply and demand.

It gets worse: a system operator's job isn't *just* to balance national supply and demand: they have many other responsibilities such as ensuring there's enough inertia on the system, and ensuring that local constraints aren't violated (e.g. for voltage). This is all driven by the exact generation mix at each location on the system. But, today, no one knows the exact generation mix at each location on the system!

Sharing data across the grid will become increasingly important as the grid becomes more complex. In a few years time, we'll look back at the grid of 2021 and think "wow; the grid was so *simple* back then!". In Britain, there are an additional 16 GW of batteries in the planning pipeline that will do their own thing; and another 10 GW of PV; and DSOs will increasingly do more local balancing; and demand will become spikier and harder to predict as transport, heating and industrial processes are electrified; and as electricity gets converted to and from hydrogen and synthetic fuels. The conventional energy companies will increasingly need help from innovative startups to help manage and predict all this! And those startups will need access to data.

To summarise: We need much better real-time information about what each part of the grid is doing. Better data will substantially reduce costs and carbon emissions.

The rest of this document proposes how the Energy Data Visibility Project fits in with the wider landscape, and helps bring us a step closer to having a "NORAD" for energy data.

Why the EDVP alpha is so important

My interpretation of the EDVP alpha is that the project is about:

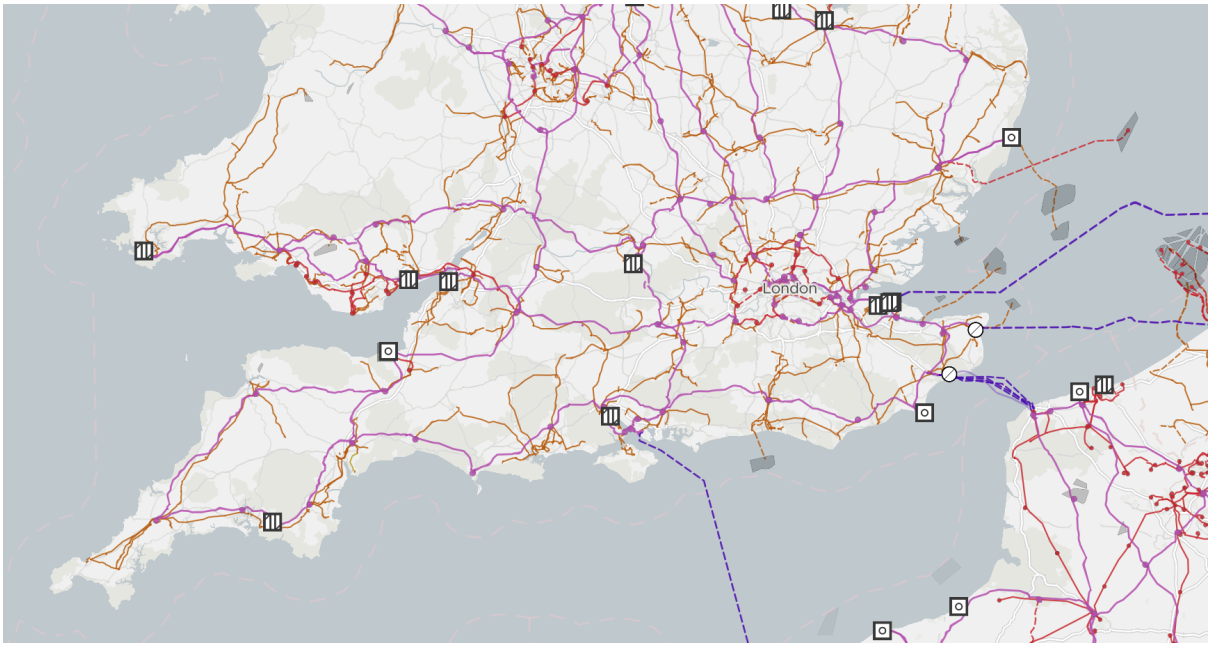
- Defining (or selecting) metadata standards for energy data.
- Converting some existing metadata into this new standard.
- Building a central representation of this distributed metadata.
- Building a front-end to easily search the metadata.
- Testing if this solution satisfies users' needs.

I think this is super-important because:

1. The opportunity isn't *just* to define a metadata standard that captures the dataset creator, date of creation etc. That's been done, and frankly isn't super-interesting. The opportunity here is to lay the foundation of a fully distributed digital representation of the entire energy system, and the datasets recorded by the energy system (as sketched out in the 'Linked Data for the Energy System' blog). A representation that's living and breathing. To take the best bits of "linked data" and apply them to the entire energy industry. To fundamentally transform the way we think about energy data, and to massively improve the chances of the wider ecosystem building innovative solutions to reduce CO2 emissions. To realise some of the things that us energy data geeks have been dreaming about for years: open energy data that's super-easy to find and consume, enabling cutting-edge research, and enabling a step-change improvement in situational awareness & energy forecasting (amongst many other use-cases).
2. Of course, the EDVP can only take the very first steps towards this goal during the EDVP alpha. But it's *essential* to get those foundations right.
3. Metadata standards stick for decades. It's unbelievably important to get this right before it's locked in; otherwise we'll be living with the consequences for the remainder of our professional lives. I don't see there being another opportunity like this in our lifetimes, I really don't! (It's also essential that the metadata can evolve over time: Perhaps the most important thing to get right from the beginning is the governance of the technical standards for the data and the metadata. More on that later…)
4. It's an alpha, so we should fully expect the *code* to be thrown away. But the *ideas* will be influential (assuming they work).
5. And the timing is great:
    a. Ofgem, BEIS, ESO and the DNOs are all super-focused on data right now.
    b. If we don't demonstrate a much more ambitious vision in 2021 then far less powerful approaches to metadata may be locked in for decades to come, much to the detriment of the entire energy system, bill payers, and our ambitions for reducing CO2 emissions to net-zero.

Design ideas for the EDVP

Energy data is intrinsically spatial



(image from OpenInfraMap.org)

All energy assets exist *somewhere*, and that location is super-important. The questions that people want to answer with energy data are often related to the spatial positions of energy assets. For example, renewable energy developers ask questions such as: *where* to install the next renewable energy project; *where* is a cheap place to connect to the grid; *where* is land available. Local authorities want a list of all energy datasets for their region. Energy forecasters want to know the location of renewable energy generators.

As such, searching for energy data feels more akin to planning a road trip (remember those?!) or searching for a local restaurant on Google Maps than searching for a new laptop on Google Web Search. (That said, the EDVP should allow users to search using *both* an interactive map and a text search like Google web search. Whilst *most* searches will benefit from a map visualisation, some searches will be a better fit for a textual web search.)

Represent both energy assets and energy datasets

Energy assets and energy datasets are two sides to the same coin, a little like musicians and songs. Assets produce data (just as musicians produce songs). Data describes assets. When you search for music on, say, Spotify, you can search for the name of a new song you just heard on the radio. Then, to find similar songs, you can click on the musician's name to find similar musicians, and then find popular songs by those musicians. Now imagine an alternative world where songs and musicians are represented in two totally separate

databases. The 'songs database' has no concept of 'musicians'; and the 'musicians' database has no concept of 'songs'. Not only is it more work to build two separate databases, but separating 'songs' and 'musicians' also makes the database far less useful. So, to return to energy: it will be less work *and* will result in a far more useful system if we build a distributed representation of *both* assets and datasets, and the relationships between them. This will also help to solve two problems at once: asset registration and data visibility.

Long-term governance of technical standards should be 'presumed open'.

To be blunt, a lot of decision-making in today's energy industry is far too slow and opaque. In contrast, the open source software community has a long history of agile, inclusive decision-making which could be adapted to the energy community. Technical decision-making in the energy industry, like data, should be 'presumed open'.

One of the great things about the energy industry is that people stay in the energy industry for many decades. But that also means that some senior folk simply don't have hands-on experience of advances in neighbouring industries. Technical decision-making would benefit hugely from a more diverse range of views: Include those who are on the frontline and have day-to-day experience of the technical issues under consideration, and include folks in neighbouring industries. For example, wouldn't it be great if computer science PhD students could help develop technical standards for the energy industry as part of their work. Or one of the many highly-knowledgeable 'energy enthusiasts' could dive in and fix a particular issue.

Technical decision-making also needs to become much faster and more responsive. Don't wait until a complete draft of a new spec is ready before running a lengthy consultation that takes a day's work to respond to, and stays open for half a year. Instead, release designs early, and let the community make comments on as much or as little of the proposed design as they wish. Evolve the design in full view. (Although only allow the core team to *accept* proposed changes.) The design process should be more like a real-time conversation. Everyone from core technical contributors to Ofgem to BEIS to interested lay people should interact with the design process in the same way. This is how hundreds of open source software projects are managed, including ones which run the vast majority of the Internet.

Several places to learn from:

- OpenStreetMap decides its schema in an entirely open and collaborative way.
- W3C standards (like DCAT) are openly discussed on Github.

Controlling access to data

A lot of energy data can be openly shared for free. This is especially true for data from deep within the energy system which describes physical infrastructure that's been paid for by everyone, and where the data doesn't reveal anything personal or commercially sensitive. It's been said before that this energy infrastructure is a little like our national parks: paid for by everyone, for the benefit of everyone, and accessible by everyone.

But there is also plenty of data which cannot be openly shared. There are two broad reasons for not openly sharing data for free:

1. **Sensitivity**: I don't want potential burglars to see when I'm on holiday by getting access to my home's real-time smart meter data! That data is private to me. But I'm happy for regulated entities to access that data if it helps the wider grid. The Icebreaker One Open Energy Directory is purpose-built for controlling access in this way. The Directory provides a digital certificate to prove identity. This digital certificate is a little like a passport. That certificate could be given to regulated organisations to prove that they are regulated by Ofgem. The dataset metadata could indicate whether an Open Energy digital certificate is required to access the underlying data. If it is, and if the user already has a certificate, then they should be able to seamlessly access the data, in the same way that you don't have to re-enter your password every time you visit Amazon.com.
2. **Money**: Some data will only be shared in exchange for payment. For example, if you buy a solar PV inverter today, chances are that it will 'phone home' to send live data to the hardware manufacturer. The manufacturers are keen to make money from all the data they collect (not because they're greedy but because profit margins are razor-thin in the solar industry, so manufacturers are desperate to claw back any profit they can). But, for solar PV data to be really useful, it needs to cover a high proportion of the solar PV systems in a country. So maybe manufacturers could club together to form a 'data co-operative' which aggregates data from multiple manufacturers, standardises the data, and anonymises the data by summing it up across multiple PV systems. Any money flowing into the data co-op will be given to the manufacturers. But, crucially, once the data is anonymised then the data can be given to anyone who has paid. The data isn't personal or commercially sensitive. Data buyers could directly purchase data from the data co-op, just as you do when you buy from any other seller on the web.

Different access controls for different views of the same data

Smart meter data from individual homes is very private. But, if you aggregate that same data across geographical areas, then you can no longer infer any individual's private behaviour from the data, so the aggregated data can be shared openly.

Live, spatially aggregated PV power data might be worth money (because grid operators and energy traders want live data). But historical data should be given away for free to encourage research. This is standard practise in the weather forecasting community, where 'live' forecasts cost money, but historical forecasts are freely available. Giving away some data also helps to advertise those datasets and so, ultimately, giving away some data should increase total revenues for the data providers.

Describing access control in the dataset metadata

The dataset metadata should be able to express these concepts in a machine-readable way. For example, the metadata should be able to make statements like:

- "This dataset of real time, spatially aggregated PV data is available from PvDataCoOp for a yearly fee. But historical data is available for free from <link to other dataset metadata, which describes the historical version of this data>"
- "This dataset of raw smart meter data is only available to people in possession of an Open Energy Directory certificate. But spatially aggregated, realtime data is available for free from <link to freely available, aggregated dataset>"

It will take years to build the full solution.

Selecting or designing technical standards for data and metadata will take years, and should not be rushed. We're laying the foundations for the digital energy system that will last for decades. It's essential to get the foundations right.

What can be achieved in the 3-month EDVP alpha?

3 months isn't enough time to build a full prototype of the entire system described above. In 3 months it'll only be possible to interpret a fraction of the data out there (because data is in a vast array of data structures). Interpret as many 'real' datasets as possible. For the demo, maybe also add some 'pretend' datasets to fully illustrate the visualisations.

According to Government Data Services, a priority for an 'alpha' is to test our hypotheses about what *users need*. We can be fairly certain that, given enough time and attention, the *technology* will work. The most important question to explore in the alpha whether the proposed solution satisfies users' needs. So we could prioritise building a functional but gorgeous map of energy datasets and the assets associated with them, using as many pre-existing technologies as possible. For example, CKAN provides a good backend for collating datasets, and then we could build a custom front-end including a map [Flo & Thought Bot's idea!].

In terms of preparing metadata, we could:

- Prepare a list of requirements for energy metadata
- Search existing metadata schemas and assess them each against this list of requirements (DCAT looks promising, although not sufficient on its own)
- Convert a number of existing metadata / data into the proposed standard…

What exactly does the metadata need to express?

Some concepts are already captured by DCAT v3:

- Temporal and spatial coverage
- Temporal and spatial resolution
- Update frequency
- Date of last update

Some concepts aren't captured by DCAT

Perhaps we can extend DCAT to Energy, and propose this as an official extension to W3C. (DCAT is designed to be extensible, and has been extended to other domains in the past).

Other existing ontologies are likely to capture most or all of these (e.g. the Open Energy Ontology; note that the Open Energy Ontology is distinct from the Icebreaker One Open Energy MEDA project)

Perhaps we can 'pick and mix' from existing ontologies, rather than writing our own from scratch. Some concepts we'd like to represent:

- What physical quantities are measured by each column? (e.g. active power flow, voltage, etc.)
    - What are the units?
- Polarity: what do positive and negative values mean? (e.g. positive means power flowing from a battery into the grid, negative values mean the battery is pulling power from the grid)
- Is this dataset an updated version of a previous dataset?
    - If so, what's changed since the last version, and where can we still find metadata about the previous version?
    - Or is the previous version unavailable?
- A list of assets described by this dataset (e.g. Electralink's data transfer service captures data from >300,000 MPAN meters. Would we list all those MPANs in the metadata?)
- What exactly do timestamps mean
    - Timezone
    - Period ending or period beginning?
    - How was the period aggregated? E.g. is each value in the dataset the mean of second by second readings recorded over the previous hour?
- Is this dataset an aggregated version of another dataset? If so, link to the upstream dataset and describe the aggregation technique.
- Access control
    - Is an Open Energy access control certificate required to access this data?
    - Is payment required? If so, what's the cost?
    - If access control is required, explain why.
    - Is a different view of this data available openly for free? (or maybe it shouldn't be the dataset metadata's responsibility to link to metadata describing different slices of the same underlying data. Instead, each slice of the data would link to the metadata describing the asset from which the data was recorded, and the knowledge graph can then list all the alternative datasets for each asset. Or, maybe take a 'belts and braces' approach and do both!)

People don't want to write metadata from scratch

Most people won't want to create raw metadata files in a text editor. So why bother creating a rich metadata standard?
A good analogy is HTML (the file format which describes every web page). That vast majority of the world's web pages were created by people who have no idea how to write HTML from scratch. If you want to write a blog, use Medium.com. If you want to share a video on the web, upload it to YouTube. And, even hard-core web developers don't write much HTML now-a-days: There are automated tools for creating HTML (e.g. blogs which convert markdown to HTML; or content management systems which dynamically generate HTML from content in a database).

But, nevertheless, it's absolutely *essential* to have HTML: It's the expressive, common language of the web. Once a common language is agreed upon, then the community can create tools to author, validate and consume information in that common language.

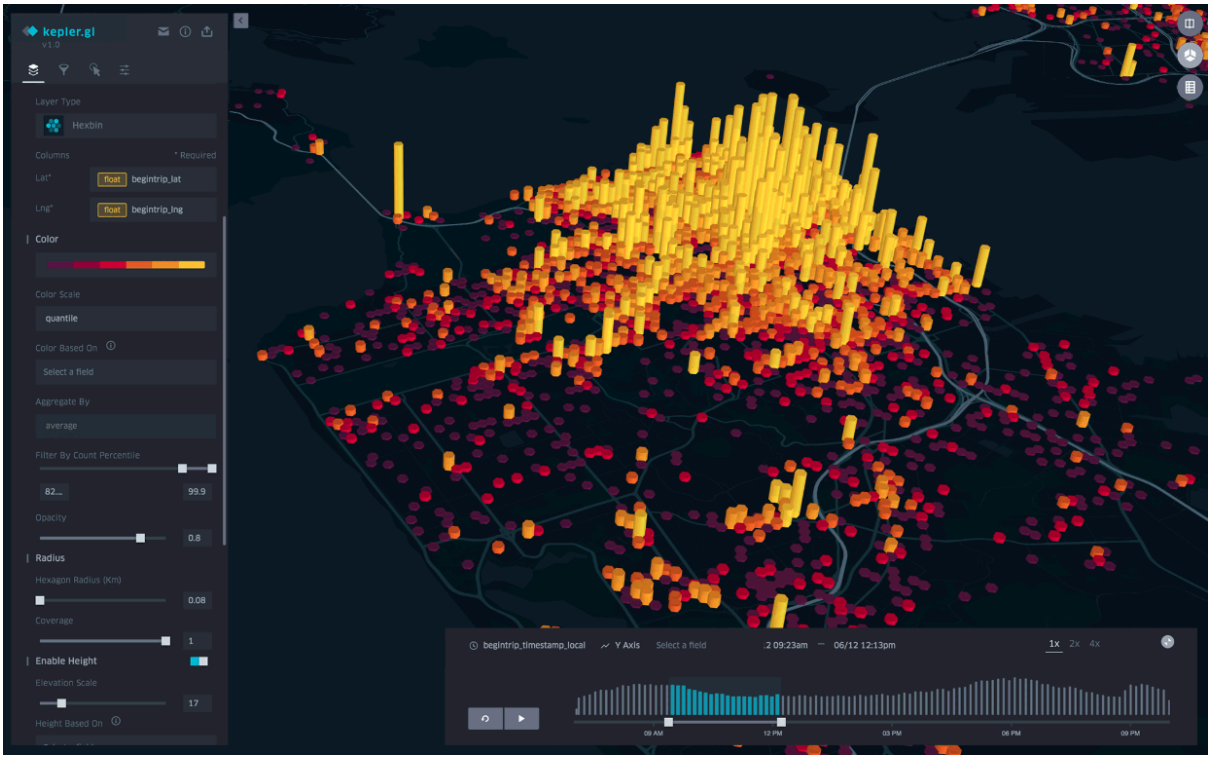
Tools for helping to create energy metadata

- A web form for creating metadata. A "wizard" which asks the user questions and outputs metadata.
- Open source scripts to convert metadata from existing formats into the new format. e.g. a script which automatically converts all the Embedded Capacity Register Excel files into the new metadata format describing energy assets.
- A web 'validator' tool. Upload your metadata files and the validator will check that your metadata is well formed. Errors will be presented in a friendly way.
- A web tool to create metadata from Excel files. Author new metadata in the Energy industry's favorite tool: Excel. Upload the Excel file and the web tool will output validated metadata.

Collaboration with industry during the EDVP alpha

- Discuss with the community in one-on-one meetings and larger webinars
- But, crucially, these are in no way sufficient to design metadata standards. Designing metadata standards requires careful consideration over long time periods. So, during the EDVP alpha, publish draft designs in Google docs / GitHub and invite everyone to comment and suggest changes; and discuss in public online discussion platforms like Twitter and the OpenMod forum. The key is to make it super-simple for busy people to comment & make suggestions. OCF has used these approaches since our formation and found them to be hugely productive (and the community is hugely generous with ideas).
- Designing metadata schemas is surprisingly hard to get right. Engage enthusiastically with "knowledge engineering" academics and practitioners.
- Work closely with the ENA & Ordnance Survey energy system mapping project as well as the National Digital Twin and the Icebreaker One Open Energy project.

Technologies



- Could probably pull together a quick (but gorgeous) demo with minimal coding using Kepler.gl and a handful of geospatial datasets.

This is entirely in line with Icebreaker One's Open Energy but is solving a different set of problems

Open Climate Fix had the huge privilege of collaborating with Icebreaker One on Modernising Energy Data Access Phase 1 and 2, and OCF implemented the alpha version of the Open Energy Search in phase 2. In general, we're huge fans of the general idea of a distributed web of energy data (as you can see from our enthusiastic blog post last year proposing the use of linked data for the energy system) and passionately want to see this vision come to life.

Icebreaker One is focused on solving one specific problem in energy data: How to control access to data that cannot be openly shared. Icebreaker One's proposed solution is to adapt the Open Banking Directory, which provides access-control certificates to prove identity (a little like a passport). Icebreaker One have said that they aren't focused on defining metadata standards themselves, instead they are looking to the wider community to define the metadata standards.

That's where we come in! We see it as absolutely essential to solve a range of technical problems which are currently stopping the community from achieving the Open Energy vision. We bump into these problems every day in our work with existing energy data, and would love the opportunity to help solve these problems. In particular, how do we:

- Represent rich information about energy assets and energy datasets, including relationships between assets and datasets. To be useful, we believe it's necessary to go far beyond Dublin Core metadata.
- Uniquely identify assets and datasets.
- Automatically interpret datasets (what does each column *mean*?)
- Join multiple datasets together.

All our work will help bring the Open Energy ecosystem to life.

Useful links

- https://bidstats.uk/tenders/2021/W08/745537633
- Detailed PDF from BEIS.

TODO on blog:

- Link in with ODI.
- Long-term governance
- Mention https://www.stardog.com/resources/knowledge-graphs-101-how-to-overcome-a-major-enterprise-liability-and-unleash-massive-potential/ and Star Dog's explorer, as a way to visualise this stuff.
- Maybe some visualisations

Text moved from early drafts of the EDVP bid

What does it mean for IT systems to be interoperable; and how do we get there?
In the 1980s, Tim Berners-Lee was getting increasingly frustrated with the pain involved in finding information stored on different computers over the network. In 1989 he sat down to fix this lack of interoperability by designing a standard document format (HTML) and a standard API (HTTP) to enable any type of computer to exchange documents over the network with any other type of computer. And so was born the world wide web! The web wasn't an immediate success: it took several years to convert legacy documentation formats

to HTML and to build the tools to create and serve HTML. The crucial point here is that the web is *all about* solving interoperability.

Energy data today is like the Internet in 1985. Today's energy data systems are not interoperable because they do not speak the same language, or even share the same concepts. The fix looks easy from a distance: we need common standards and vocabularies. But, in practice, there is a lot of work ahead. The EDVP and the metadata standardisation is an essential piece of the puzzle.

Ultimate vision

We propose following the footsteps of the biomedical community, who have *excelled* at capturing relationships between vast datasets by building a distributed 'graph'. This approach was first proposed by Sir Tim Berners-Lee. (Here, we use the term 'graph' to mean a collection of entities (datasets, companies, etc.) and the *relationships between* those entities. This is distinct from the more common use of the word 'graph' to mean a data visualisation.)

We're confident this vision is achievable by taking small, iterative steps, starting with the EDVP Alpha.

Ultimate vision

To understand our approach to the EDVP Alpha, it is important to understand our ultimate vision for the EDVP production service (*after* the alpha), and why we are so excited about the EDVP!

The ultimate vision is to build a distributed, machine-readable, digital representation of not just datasets but of the *entire* energy system. A representation that's living and breathing. We arrive here by adapting a simple technical approach developed by Sir Tim Berners-Lee called "linked data" to implement one of the key recommendations of the Hippo Digital report:

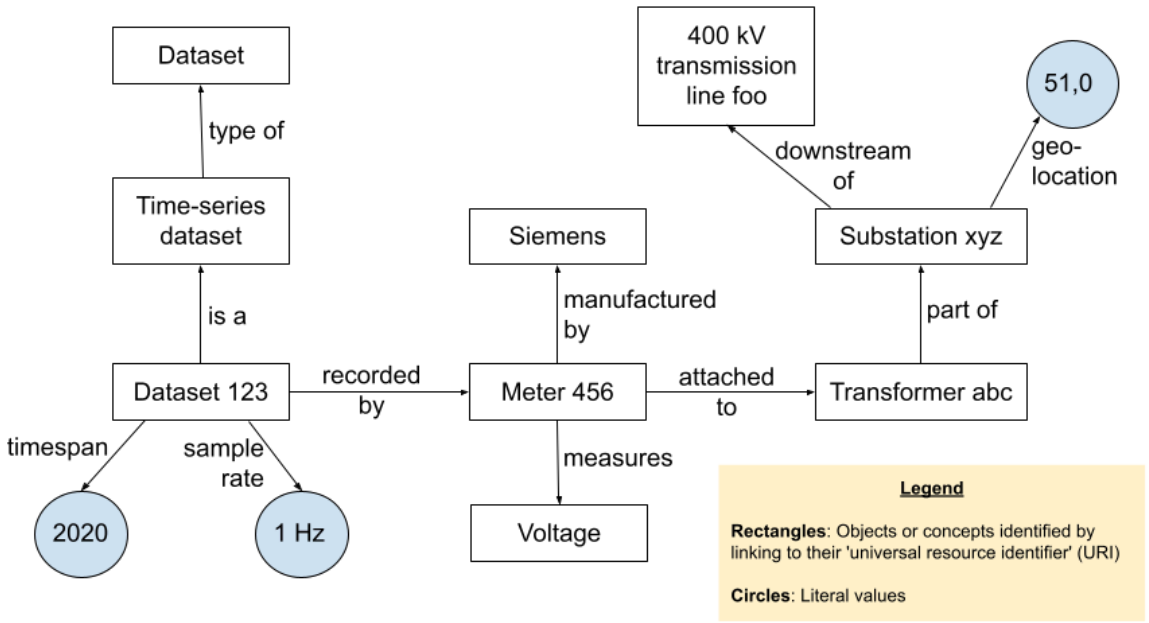
> *"In making data discoverable, searchable, and understandable through standardised attributes, the long-term goal for the service will be to support the creation of an ontology for energy sector data, wherein rich and semantic relationships between existing and future data helps unlock the underlying value across the datasets."*

Users want to search for datasets using queries like "*show me all datasets describing generator xyz*"; or "*show me all datasets released by company abc".* That is, users want to search for datasets associated with specific *entities* (companies, assets, concepts, etc.). So, the metadata must describe the *entities* associated with each dataset. But - for multiple reasons - it is bad practice to copy-and-paste a full description of every entity associated with each dataset into its metadata file. It is far better to describe each entity only *once* (and to do so using machine-readable metadata, in a standard form, and using a standard vocabulary). Then, to associate a dataset with an entity, the dataset metadata would *link* to

the description of that entity (just as web pages link to other web pages). But, to do that, we need a digital representation of the energy system *itself* (so the dataset metadata has something to link to!)

The ultimate aim is to radically improve visibility of energy data and the chances of the wider ecosystem building innovative solutions to reduce CO2 emissions to net-zero (which is what we're most excited about). In so doing, this solution will help solve three problems in one go: energy data visibility; asset registration; and the energy portion of the UK National Digital Twin.

To summarise our vision: Every entity (dataset, generator, transmission line, DNO, etc.) will be described by a metadata file in a standard form. Relationships between entities will be expressed by *links* between metadata files. The figure below gives an example of how "Dataset 123" (on the left) is linked with its associated entities:



In mathematics and computing, we call this collection of entities and their relationships a "graph" (as distinct from the more common use of the word "graph" meaning a data visualisation). The EDVP service will create a local representation of the metadata in a graph database, which will allow users to search for data from one place.

We're confident this vision is achievable by taking small, iterative steps (starting with the EDVP Alpha). Furthermore, this approach is well established: The biomedical community makes extensive use of exactly this approach to build a distributed, linked representation of vast quantities of biomedical data. And the approach described above is inspired by - and entirely consistent with - the UK National Digital Twin. And we do not need to start from scratch: There are existing metadata standards for the energy system that we can combine and build on.

We will make life as easy as possible for dataset publishers by building tools to make it easy to create, validate and consume metadata; we will create converters to convert existing

metadata into the new draft format; and we will help the community collaboratively develop additional converters.

Of course, we cannot build this within a 3-month alpha! During the EDVP Alpha we will answer users' most pressing needs by implementing a simple graph database, and we will begin collaboratively developing draft metadata standards with the community. During the EDVP Alpha we will test our riskiest assumptions.

International projects & standards
There are many relevant international projects that the consortium intends to build on. These include the Open Energy Ontology (not related to the MEDA Open Energy project!); Schema.org; the Common Information Model (CIM - which the Linux Foundation Energy project are pushing to make partially open-source); the Data Catalog Vocabulary (DCAT - which is already used by many energy data catalogues & is a good candidate 'foundation' vocabulary for the EDVP metadata); Web of Things (WoT); the Scalable Policy-aware Linked Data Architecture for Privacy (SPECIAL); SzenarienDB; Spine; DBpedia Databus; PowerSystems.jl; the Public Utilities Data Liberation Project (PUDL); Open Power System Data (OPSD); the Pangeo Cloud Datastore; Open Infrastructure Map (which visualises data from OpenStreetMap); Semantic Web; Linked Data; Resource Description Framework (RDF - the foundation for Schema.org, WoT, OEO, CIM & DCAT); Web Ontology Language (OWL); and JSON-LD (where 'LD' stands for 'linked data'). Robbie Morrison has also kindly shared with us a preprint of a highly relevant paper titled "Advancing FAIR metadata standards for low carbon energy research" which is a write-up of a 2020 EERA data workshop.

Roadmap

Although these features will not be part of the MVP or developed MVP, these features will be crucial to the long-term success of the data catalogue as we move to the beta and into full production. We have provided details of them here to create confidence in the long-term vision of the data catalogue.
Advanced Metadata modelling
Other sectors such as the biomedical community have had success in sharing structured data using full ontologies. We propose capturing energy metadata using distributed, machine-readable metadata files, written using a standard vocabulary and using a standard format. These files will be automatically stored in the EDVP graph database to be presented in the data catalogue interface. We will develop tools and templates to help create and validate metadata (just as tools and templates exist today to help create HTML) and make life as easy as possible for dataset publishers. The metadata standard will be owned by and evolved by the community, out in the open. In contrast to the Discovery report, we believe it is necessary to lay the *foundations* for a full ontology during the alpha because it will be extremely hard to change the foundation once the community starts to adopt the standard. We are confident that we can have the best of both worlds, we can use a *simple* ontology during the alpha (to rapidly test our hypotheses) which can then be extended in the future. We envision the community collaborating to write a library of scripts to convert legacy

metadata into the new format. These converted metadata files will be public and available for consumption by the rest of the community, such as the MEDA Open Energy Data Search.

Feature request

An essential part of the success of a system such as the data catalogue is its ability to adapt to new user needs. In the long term we will build a robust feature request system which will allow technical users to report bugs, suggest changes, raise new feature requests and offer feedback on how the data catalogue is running. These requests will be sourced from the energy data community to reflect the needs of the core users.

Change Management and consensus

The energy sector has extremely robust change management systems in place which allow for collection of feedback and the consideration of consensus in decision making between multiple stakeholders and groups of stakeholders. In order to ensure that the data catalogue can keep up to date with the latest changes, we propose a decentralised consensus system that will allow the energy data community to rapidly adapt to changing requirements for their data and metadata. Once metadata standards are starting to emerge, we will leverage the OE governance platform to establish the metadata standard owners. Users will be able to raise change requests to the standards and templates and the community will be able to vote on whether to accept or reject the changes and justify why they have done so. This will also allow decisions to be made much more rapidly as they can be taken asynchronously by many users.

There are eight main questions regarding metadata that we plan to discuss with the community:

1. **How should the metadata standard be governed over the long term?** The EDV Discovery report found that data providers want "*an approach to standards, architecture, and governance that doesn't present significant operational or financial overhead*." We will lay the right foundations for metadata governance during the EDVP Alpha because the metadata will evolve over the coming years and decades as the energy system evolves. As pointed out in the Discovery report, the governance must be as easy as possible. Our proposal (for discussion with the community) is that metadata governance should be based on blending UK energy governance approaches (such as DCUSA) with established, open, collaborative approaches for evolving metadata standards such as DCAT. We will use community feedback to address issues such as evolving metadata standards and addressing inclusivity of stakeholders in industry, academia and the National Digital Twin. Some specific questions to discuss with the community include: How do we ensure that the standard can rapidly evolve in response to the industry's needs, whilst not becoming too unwieldy, and not introducing backwards-incompatible changes? How do we ensure technical decision-making is as inclusive as possible, whereby we can include experts from neighbouring industries, the National Digital Twin, academics, etc.
2.
3. **How do users want to search the EDVP for datasets?** Which categories of search would be best done using a graphical interface? Would a searchable map be useful for some types of search?
4.

5. **What are the industry's *requirements* for a metadata standard?** What *concepts* does the metadata standard need to represent? We will use the list of attributes in the Hippo Digital report "recommendation 2" as a starting point, although we are confident this list needs to be extended with, for example, an attribute describing the geographical coverage of a dataset. This discussion will also explore the 'glossary' recommended by the Hippo Digital report and whether it is appropriate to use a 'proper' ontology whereby most entities, concepts and relationships are identified by universal resource identifiers.

6. **Which metadata standards currently exist?** As well as asking the community, we will also conduct our own desk-based research to discover metadata standards.

7. **Do any existing standards fulfill all the requirements?**

8. **If not, how should we extend and/or combine existing standards?**

9. **What *tools* are required to help create, validate & consume metadata?**

10. **How are other countries solving this problem?** Can we work with international colleagues to define standards that can work globally?

    The consortium will seek to achieve consensus in the sector on the answers to these questions by using the approaches described above. In addition we will use the following decision making proce