Appendix 1b

Data Strategy 2023-2028

Cambridgeshire County Council

Version: **1.0**

Date: **21st June 2023**

1 Vision

To value data as a core asset, curating it into high quality data 'products' enabling Insight and intelligence to be accessible to anyone that needs it, underpinning the design of every council service and informing every council decision. A data centric culture will flourish across the organisation enhancing personalisation, prevention, automation and innovation in service delivery while being mindful of information governance, ethics and cyber security.

2 What do we mean by Data?

Data can be described as raw and unprocessed facts that are captured for an intended purpose. For example data could be numbers in a spreadsheet, the text of case files in a database or media files for identity verification. It can be recorded and stored in digital forms such as in Business Systems or in physical form such as paper records.

Organising, categorising, calculating, and providing context to the data such as relating it to people, place and time, gives it real meaning and the result can be termed as **information.** This information is further enhanced into **knowledge** by interlinking and associating information together providing further context and understanding based on relationships, comparisons and experience. The true value of data we record and hold can be unlocked when the data is transformed throughout the organisation into knowledge that will provide us with **insights** and **intelligence**.

Data can broadly be categorised as either structured or unstructured or somewhere in between.

Structured data is formatted and modelled to fit a set structure when it is stored and the most common example of such is data that resides in relational databases with the structure being the database schema. Schemas are designed to make storage, search, analysis, segmentation and extraction of data as efficient as possible. Structured data is mostly quantitative, made up of objective well-defined facts and numbers. This strategy primarily focuses on structured data; to facilitate exploitation of data residing in business systems across the council and related external data sources.

Unstructured data is data which has not been processed into such pre-defined structures or models and in our context this most commonly resides as either textual formats at source e.g. documents, emails, social media posts; or non-textual formats at source e.g. video, audio, images. Unstructured data can have an internal structure such as with Excel files, but it's not predefined through a set data model at source, hence the distinction. It might be human generated e.g email correspondence, business documents or machine generated e.g. IoT sensor/telemetry data, CCTV footage. Unstructured data is often qualitative, containing subjective representations and opinions.

Some data can be considered between structured and unstructured, described as semi-structured data. It has some consistent and definite characteristics but is not in a rigid structure suitable enough for relational databases. It does however contain properties like metadata or semantic tags which are used to make the data more manageable but this data can also be variable and inconsistent. Examples are CSV, XML, JSON, HTML. Much of what is often classed as unstructured data, is actually semi-structured, because it contains some classifying characteristics. Section 10 will provide a brief overview of the aspirations for managing unstructured data going forward.

It is important to note the limitations of data, and the importance of qualitative insights, lived experiences, historic and place-based knowledge which may not be represented quantitively but are fundamental to understanding and interpreting issues of importance. Insights from data alone are limited to the data that is available and potentially can be misrepresentative due to the data that is not available. Questions and issues will always emerge where existing data is insufficient and there will be important considerations about what data does not provide us with and what methods can be used to mitigate deficiencies in data, especially deficiencies in qualitative data.

3 Why this matters

Understanding, improving, and harnessing data effectively into insight and intelligence supports performance management and efficiencies in operational processes, informs better and more robust decision making and empowers our citizens and staff across all aspects of council service design and delivery.

This doesn't happen automatically; we need to carefully curate and develop data across all areas of our business and treat it like an asset so we can use it to help achieve our organisational objectives. There are areas of the business in particular where we know that unlocking and exploiting data is paramount to overcoming challenges and pressures the council faces now and over the next few years. The following examples illustrate just some of these opportunities:

a) Providing intelligence to front line workers supporting individuals/families
Linking various datasets, internal and external, e.g social care, education, health, police, housing, so that care and support professionals are given access to a wider range of relevant information about families or individuals they are supporting. This would allow care and support professionals to holistically and effectively assess individuals or even whole families significantly reducing the amount of time and resource required while also speeding up decision making and altogether improving important services such as safeguarding.

b) Performance management and understanding the impact and outcomes from services delivered
Analysing the data associated with delivered services such as how many citizens are using various services, the associated key performance indicators (KPIs) and satisfaction levels etc enables continuous improvement and better targeting of council resources. Also, assessment of the overall impact of individual services and support provided at various points during service delivery ensures improvements can be made based on outcomes. e.g. knowing and tracking families/individuals that have been supported to understand what impact various interventions have had.

c) Enabling needs led commissioning of support at varying geospatial levels or for different groups
Analytics can be used to 'cut' data by demographic characteristics or at various spatial levels using GIS (Geographic Information Systems) tools to enable services and support to be designed and commissioned in response e.g. based on school catchment areas targeting particular schools for interventions such as tackling obesity or violence; or based on ward/parish levels, working with third sector organisation to help shape community and voluntary support as another example.

d) Developing risk models to assess and identify likelihood of demand for services and enhance preventative strategies
Individuals and families requiring support are likely to share some common risk factors. Analysing data to identify relevant risk factors can then be used to generate risk modelling to identify those who have not yet experienced negative outcomes but have some or several common risk factors - thus allowing early planning of support provision and preventative interventions e.g. identifying older adults who are likely to need care within two years. This modelling could be enhanced with Machine Learning and Artificial Intelligence capability.

e) Enabling local authority responses to emergencies, better protecting our residents.
Emergency planning and response is reliant on the ability to manage and share data quickly and effectively. The most recent example is of course the local authority response to COVID-19 where authorities rapidly had to take on responsibilities for supporting residents throughout lockdown and those who had to isolate as cases or contacts. This work required considerable data management capacity and infrastructure, as well as working closely with partners in the district authorities and with NHS colleagues, sharing more data than ever before.

f) Visibility and scrutiny of holistic spending on contracts.

Linking up ERP transactions to Line of Business System records to contract information on common identifiers will provide better visibility of holistic spend data with suppliers to expose spending in departmental silos and improve ability for internal scrutiny. More accurate measurement of the cost of services and Cost Per Transaction, enabling smarter decisions requiring less effort and resources. Improve transparency for citizens and politicians thus improving dialogue, engagement, reputation, and trust.

g) Property portfolio utilisation

Joining up council property asset data with data on how employees, citizens and partner organisations are interacting with property assets can provide more intelligent decision making capability concerning the future of the property portfolio of the council. For example, do the geographic locations of our properties meaningfully relate to where the demand for our services reside, how do they relate to where employees live and what is the environmental impact of this. Questions like this are fundamental to how the councils will operate in a post COVID era where patterns of working and methods of service consumption have drastically changed.

4 Progress so far

Over the last three years the importance of advancing data and analytics capabilities have been treated as a priority across the authority and good progress has been made in this area accelerated partly, no doubt, by data demands driven by the COVID-19 response. Data engineering functions providing data preparation, curation, extraction, transformation and loading, have been well established supporting many areas of the organisation culminating in a breadth of knowledge and experience in the diverse local authority data portfolio specifically amongst centralised teams. These capabilities have been built using tools and technologies which had been geared primarily around an on-premises Business Systems infrastructure. From the production of complex data warehouses and development of data matching algorithms and rules engines to the rollout of Power BI Premium for advanced visualisations and dashboards, the experience around data to date has been varied and sophisticated but requires modernisation and evolution to a cloud-first focused model and data skills and culture needs further dissemination more widely across the organisation outside of centralised teams.

The last twelve months have also seen progress in implementing industry-standard best practices required for establishing high-performing database DevOps teams, delivering dedicated development environments, version control, continuous integration and repeatable deployment models, championed by the shared IT & Digital Service. A basic data platform architecture exists with storage and Extract Transform Load (ETL) capability, consisting mainly of on-premises SQL Server and SQL Server Integration Services (SSIS) with some limited use of cloud-based Dataflows (hosted in Azure Date Lake) via the shared (CCC and PCC) Power BI Premium implementation. There also remains limited use of on-premises hosted SAP Business Objects, mainly used as a semantic layer for the Liquid Logic system (Supporting Children's Service Reporting) and for ad-hoc querying by a small number of Policy and Insight team developers (formerly named Business Intelligence team). Azure DevOps adoption has provided cloud-based repositories and Continuous Integration/Continuous Deployment (CI/CD) capabilities, creating real efficiencies in data operations (DataOps).

Recent PCC data engineering projects have utilised CCC infrastructure and resources successfully and there is recognition that a levelling-up between the organisations is necessary, particularly when it comes to the production of fast, automated data pipelines that can eliminate the requirement of manual effort and cumbersome human curation. However, going forward this levelling up needs to be done, not entirely with the existing infrastructure that is currently in use at CCC, but rather using more modern tools and technologies in line with the vision laid out in this strategy.

Predominantly as a visualisation tool, Power BI Premium has been successfully rolled out in priority areas such as Adult and Children's Social Care following an accelerated implementation during the COVID-19 pandemic. The technology has provided access to improved dashboarding and reporting capability and enabled secure data sharing with external partners and organisations. These capabilities have been sufficient to replace Qlik Sense in PCC, reducing the reliance on on-premises hosted SQL Server Reporting Services (SSRS), although SSRS and Business Objects continue to be used for reporting in both CCC & PCC, albeit without the same levels of governance, standardisation or control as Power BI. Production of information using these tools is mainly via specialised central teams, but a level of self-service has been introduced for access requests to content residing on Power BI and the granting of that access.

Extensive work has been carried out in CCC to improve computational compliance and data lifecycle management, ensuring legacy data is retained, archived and deleted at the appropriate time in a more intelligent and more automated fashion; a fully configurable in-house developed solution went live for this in April 2022. It comprises of a matching engine (currently matching service users/clients across seven legacy applications) and a business rules engine to implement data governance rules while enabling secure report lookup and querying.

In terms of Geographic Information Systems (GIS), CCC predominately uses desktop based MapInfo as the corporate GIS platform with network file shares used for data distribution and storage. Intranet and internet spatial data mapping is provided by Astun Technology's iShare In The Cloud platform. A hosted spatial data warehouse to replace network storage has been implemented on the same platform and is currently in the testing phase.

5 Pillars of Effective Data Use

We have set out the following interconnected pillars drawn from key parts of the National Data Strategy representing areas where our efforts need particular focus to progress towards the strategic vision of a data driven council.

5.1 Data Foundations

Ensuring data is fit for purpose. Maintaining the quality and integrity of data by establishing standards, processes, policies, governance and adopting best practices related to all aspects of data management.

- Assessing and monitoring data quality across all council systems on an ongoing basis
- Adopting and implementing consistent data standards across the council in terms of collection, storage, maintenance, analysis, and interoperability of data, ensuring industry best practices and common open standards are preferred over proprietary models where possible. APIs are and will continue to be fundamental to this to facilitate the secure exchange of data.
- Ensure Data redundancy is reduced by assessing and minimising duplication of data across the council and strive towards establishing canonical models where we establish single source of the truth (SSOT) for data items and entities.
- Continue to streamline data lifecycle management to ensure data is retained, archived and deleted at the appropriate times in a more intelligent and automated fashion.
- Ensuring data we hold on behalf of partners is held and used legally, using appropriate methods for exchange, handling and storage.
- Ensuring data we share with other partners is done so under a clear legal basis with a supporting data sharing agreement.

Key Deliverables:

1. We will adopt and apply **The Government Data Quality Framework**[1] which defines principles for effective data quality management and provides guidance on practical tools and techniques which can be applied to assess, communicate and improve data quality.
2. We will assess and optimise the council's 'data maturity' using **The Data Maturity Assessment for Government**[2]. This framework is used to analyse stages of progress along the journey towards data maturity. The framework tackles ten important topics within the data ecosystem covering the following:
   - Engaging with other organisations/partners
   - Having the right data and analysis skills and knowledge
   - Having the right systems in place
   - Knowing the data we have
   - Making decisions with data
   - Managing and using data ethically
   - Managing data operations
   - Protecting data
   - Setting our data direction
   - Taking responsibility for data
3. We will set up a multidisciplinary team as an internal **Data Standards Authority** who will seek to support services in identifying and improving data standards, guidance and best practices in the management of data across the council, operating and making decisions in a similar fashion to a Technical Design Authority and collaborating with other local authorities while drawing on advice and guidance published by the government's Data Standards Authority[3] where applicable.
4. Producing further guidelines for incorporating and analysing qualitative insights
5. Enhancing our technical requirements, data standards and API specifications for all new systems procured or developed, to ensure we are adopting technology and systems which are in line with our ambitions.

5.2 Data Discovery

Being in control of our data. Discovering, understanding, cataloguing, classifying and labelling data to ensure we are aware of what data we hold and where, why we hold it, how it was derived (provenance) and the potential value of the data.

- Defining clear governance and ownership of information assets.
- Anyone involved in design, commissioning, and management of services will need to be more accountable for the data associated with their services.
- Describing key information about the data we hold to establish rich meta data
- Ensuring data privacy, regulatory/statutory compliance (e.g. UK GDPR) and ethics are always covered as priority. Privacy by Design will be fundamental everything we do.
- Producing simpler data access and sharing agreements to promote the flow of data within the council and sharing with partners externally.
- Working across the organisation to exploit the potential insights that can be gained from the data through intelligent and appropriate analysis.
- Reviewing consent models for citizen data sharing and ownership. How can we empower citizens to own the data we hold on them.

[1] https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework
[2] https://www.gov.uk/government/collections/data-maturity-assessment-for-government
[3] https://www.gov.uk/government/groups/data-standards-authority

Key Deliverables:

1. We will define key capabilities and best practices using a **Data Governance Framework** and ensure this is embedded across the council.
2. We will build an enterprise-wide **Data Catalogue** to act as an inventory for all the data held by the council, detailing the various information assets and their relationships and dependencies, where they are stored, ownership, accountability and for what purposes they can be used and shared as well as other relevant metadata. It will cover details of relationships and linkages with external datasets and details of how external parties are to use our datasets with gatekeeping to ensure safe and competent use.
3. We will establish an **Enterprise Business Glossary** so that everyone in the organisation understands data related terms and removes ambiguity around data items and entities and their relationships.

5.3 Data Democratisation

Making data available and useful to everyone irrespective of their technical know-how. Ensuring Data becomes the fabric of our organisation, making it available and accessible to the people who need it, when they need it, in formats which provide insight and intelligence.

- Providing the ability for staff to take insight and intelligence from data intuitively to help improve council services
- Breaking down information silos and creating greater searchability around our organisational data.
- Eliminating time wasted and advocating lean principles in data preparation.
- Defining links between data rather than just links between systems
- Making data central to the design of digital services to enhance personalisation and user experience.
- Ensuring more council data is made open fostering re-use of our data by citizens, external organisations and even within our organisation.
- Continuing to develop relationships to share data effectively with partner organisations such as NHS, Police, Universities and Care Homes
- Members of the public can find information and advice they need easily
- Members of the public can access the data we hold on them easily

Key Deliverables:

1. We shall establish a **Data Centre of Excellence** with a cross-section of skillsets as well as service based **Data Champions** (including Information Asset Owners) to disseminate skills and knowledge amongst staff as well promote and support the use of data and insight across the council and measure the impact of data initiatives. Working with and learning from other local authorities and the wider public sector and local Health and Care system.
2. Further automate and enable self-service user administration for processes surrounding access requests to datasets and their approval.
3. We shall continue to develop and roll-out our Microsoft Power BI Premium capability to ensure the ability to analyse and visualise data is ubiquitous across the organisation.
4. We will continue our commitment to accessible data, enhancing the Cambridgeshire Insight platform, including supporting publication of open data (see section 9) as well as arranging data science related code competitions and hackathons to help foster innovation in public services using open data.

6 Conceptual Architecture: Data Mesh

Establishing the approach needed to build an integrated yet decentralised data platform to facilitate a shift from 'data as a by-product' to 'data as a product'. Comprising of cloud-based data services bringing consistency by building on data foundations, enabling discovery and democratisation.

One of the challenges for local authorities when it comes to data management is the decentralised and diverse nature of the services being delivered across the council which naturally inhibits the successful implementation of traditional totally centralised and monolithic data platforms which have been a common ambition across the public sector in the attempt to break down data silos. Furthermore, our cloud-first Digital Strategy shifts us away from having data hosted primarily within our on-premises network as we find ourselves in a new reality where our data is becoming more and more dispersed across a multitude of business system suppliers and their hosting arrangements as we move more services into the cloud. Our data is becoming more scattered across the internet in hybrid and multi-cloud environments. This calls for a different approach.

A relatively new paradigm, the Data Mesh[4], devised by Zhamak Dehghani in 2019 provides a more decentralised approach which lends itself to a more distributed data landscape where services retain more control over their data domains, facilitating a domain-driven governance model where data ownership and organisation stays with the domain experts, i.e. the different teams delivering vastly different services, giving more focus to the business outcome and encouraging more self-service in various aspects of data management.

Our strategy for data architecture across the council will largely be based around this Data Mesh concept which fundamentally shifts the focus away from a traditional 'push and ingest' architecture to a more federated 'serve and pull' architecture.

This decentralised approach also supports the need for a data focused culture to permeate across the organisation instead of being centred exclusively around IT, Policy and Insight and other specialised teams such as Public Health Intelligence. It promotes and requires much development of data literacy and skills organisation wide. This does not however, absolve the need for centralised data expertise, rather existing centralised experts become more focused on what they do best and become a 'centre of excellence' in establishing, maintaining and growing the Data Mesh and the people, processes and technology which underpin this capability. As we progress on this data journey we will need to find the right balance between centralisation and decentralisation but treating data as a 'product' in its own right, rather than data just being a by-product of service delivery, is fundamental to this approach and to achieving the vision.

The Data Mesh paradigm is founded upon four core principles, namely data as a product, domain ownership, self-serve data platform, and federated governance. The 'domains' in our case are the business functions, departments and services and the four principles can be summarised as follows:

Data as a Product:
Data Mesh brings 'product thinking', and 'domain driven design' practices already used commonly in software development to data management.

- Business domains are producers of data 'products' which must be discoverable, secure, explorable, understandable (documented), trustworthy and crucially must meet the needs of the data consumers, just as any products are designed for their customers. User experience should be at the heart of data product design.
- Data Mesh defines the role of 'domain data product owners' responsible for the production and publishing of data as 'products' with the characteristics mentioned above. Although data owners already exist across the council, the definition of 'domain data product owners' is different and wider based on this paradigm and in our case need not just be one person but would be a multidisciplinary team of people akin to a product management team.
- Such cross-functional product management teams should be concerned about whether end users (consumers) are getting value from the data 'products' they are delivering and measuring their success using metrics on aspects such as lead time to data consumption and data quality. The product approach is about bringing together stakeholders, users, and experts to consider what is valuable and how we can

[4] https://martinfowler.com/articles/data-monolith-to-mesh.html

achieve it; taking an iterative approach to product design; thinking about how all the components of the data product work together and remain sustainable.

Domain oriented, decentralised data ownership and architecture:
Decentralisation and distribution of responsibility to those who are closest to the data to support continuous change and scalability.

- The benefit is that the domain's familiarity and real experience with the data will provide deeper insight into where, why, and how it should be used. For example, those in the practice of Social Care delivery will understand all the nuances of the data they process better than any central team who are looking at the data alone somewhat removed from the day-to-day work of practitioners and care workers.
- Many data entities are best generated and described by the operational systems that sit at the point of origin and so the closer related the operation systems are to the data 'products', the more detail is retained in an exploitable manner. This proximity also means cleaning and transformation is handled closer to the data source which often results in data entities being more reusable.
- Continuous change and scalability is facilitated as both responsibility and processes around data is limited to domains while data products are also loosely-coupled in terms of the technical interoperability with other data assets and infrastructure, making individual domain's product portfolios far less complex than a monolith. This results in more agility for driving and managing change, e.g. growth of data sources and demands from data consumers for each domain is managed and prioritised separately. Individual data products will have their own prioritised 'product backlogs' as agile development roadmaps.

Self-serve Data Infrastructure as a Platform:
Data Mesh requires data to be made available in a simple and easily consumable self-serve manner for analysis and insight.

- Facilitates data discovery and enables data democratisation so that data can be published by producers and is accessible to the consumers who are authorised to make use of it.
- This is the shared technology and tools used by domains.
- Our core technology for the self-service consumer end of this 'platform' is Microsoft Power BI Premium, which is already in use across the council, however much work is required to further build and establish end-to-end data pipelines that integrate, transform, and serve data in a manner where the data is sufficiently curated to facilitate effective business intelligence outcomes ubiquitously. Roles and team boundaries will change over time as a result of the technological disruption that such self-service platforms can provide.

Federated computational governance:
Modern monolithic data-lake based initiatives have often turned into 'data swamps' due to lack of robust organisation, governance, and accessibility of the data ingested into one store. Data Mesh encourages a federated governance model based on decentralisation with governance policies for each decentralised domain defined, while ensuring all domain data owners and teams operate within a consistent standard governance framework.

- Data products produced by the different domains are required to interoperate with each other and can be combined to solve new problems as they arise.
- Underpinned by adoption of common data standards and APIs across systems to ensure interoperability, globalised security and compliance.
- Policies should be managed globally; should be computable in code and configurable as rules where possible, to support automation and for data products to be easily consumed.

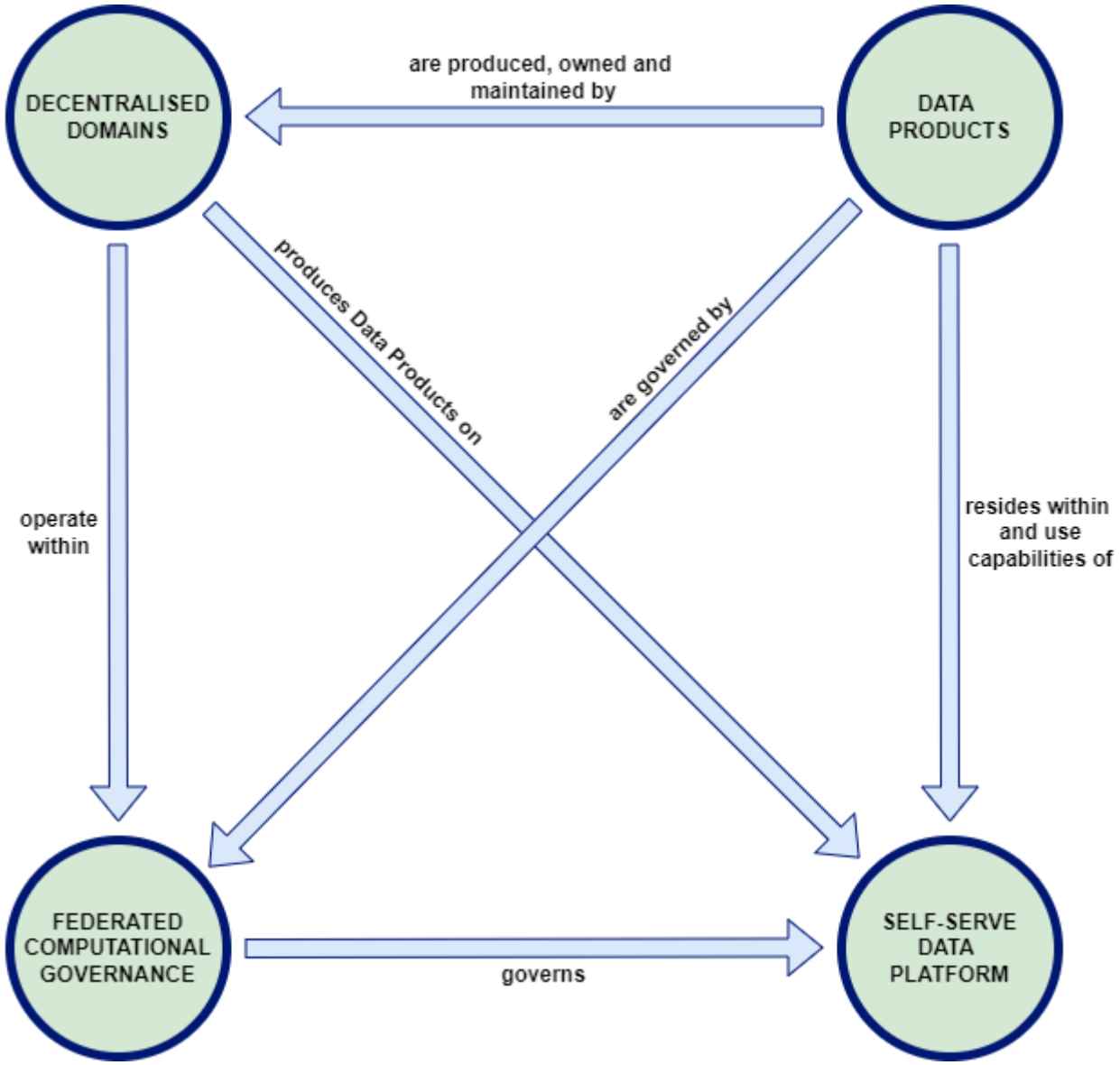

Figure 1: Data Mesh Principles

Data Mesh is more of a philosophy than a technology or toolset, and to work towards it will be an evolutionary process with various functional components implemented over time in an iterative and agile manner. The new strategic model aligns with some aspects of day-to-day practice already and the adoption of such a paradigm seeks to build upon those practices in a consistent matter across the organisation. For the very reasons mentioned earlier, in terms of diversity of use cases across the council, the nature of our data architecture going forward will be somewhat of a best of breed initiative. The Data Mesh will be composed of data warehouses and data-lakes, and its platform underpinned by data 'lakehouses' but they will not be monolithic and all-encompassing but rather in-line with the Mesh philosophy.

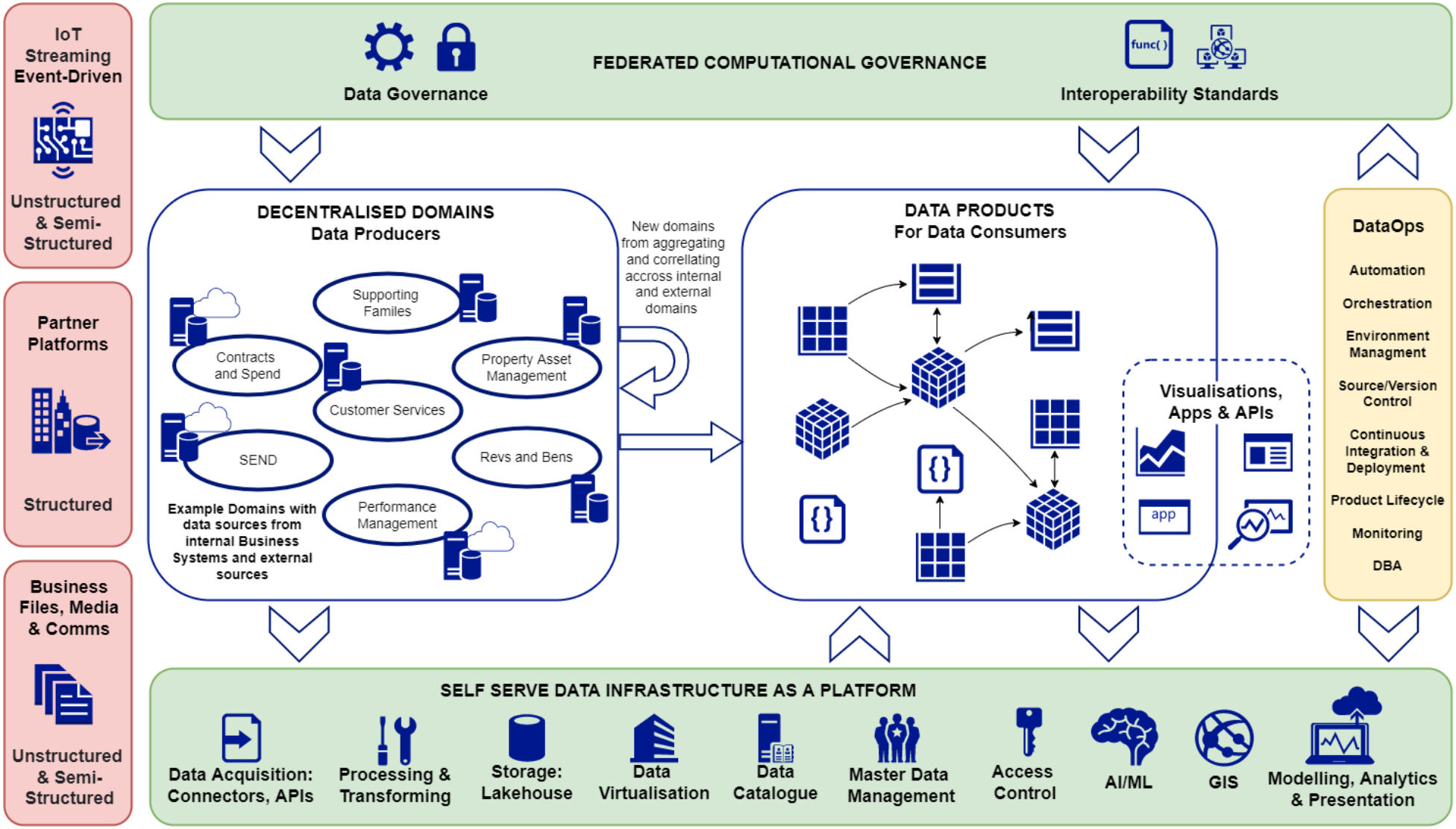

Figure 2: Data Mesh Architecture High Level Overview

7 Foundational Data Platform

Our current position on this data journey has seen much progress but lacks key data capabilities that are needed to establish a strong foothold on this renewed more modern vision for data. The practical approach to realise these missing capabilities will be to begin by implementing a Minimal Viable Product (MVP) relating to a real use case which results in an architecture encompassing many of the components that we need to establish for the long-term wider council vison for data architecture, thus providing a foundational 'self serve data infrastructure as a platform' from which these component capabilities can then be built upon and replicated for other use cases across the council. The learning from the MVP can be replicated acting as a lighthouse project for this strategy. The MVP and the deliverables pertaining to the pillars of effective data use (section 5) are not mutually exclusive however; as the continued work on data foundations, discovery and democratisation are key to scaling our data capabilities and achieving the desired vision.

This targeted first phase of the data strategy will see the creation of a data platform scoped specifically to enhance our Supporting Families programme, improving the lives of vulnerable children and families in Cambridgeshire and Peterborough as well as providing tangible savings in this area for both CCC and PCC. A product team will be established, working in an agile manner, to produce data 'products' and deliver end-to-end data pipeline capabilities for business intelligence, drawing in data from internal and external sources. The platform will be secure, scalable, reusable and be built with cloud-based data services that will need to be adopted, together with further exploitation of technology already in place such as Power BI Premium in particular. The platform components and data products will be well documented from both an end-user perspective and technical perspective in the form of 'playbooks' and 'runbooks' and supported in a DevOps like manner as DataOps, to be able to sustain new aspects of data operations in terms of support, maintenance, management and administration of new capabilities.

The development of various new data capabilities within the council may seem superficially like duplicated effort, where there are external data platforms which are being harnessed as part of partnership initiatives (see 'Working in Partnership' section 8 below). However, this is actually mitigated by the super agile nature of cloud-computing and the data technologies that are currently available, i.e. the on-demand, rapid provisioning, flexibility of the technology and the consumption-based pricing models. The people aspects, skills and processes that come with building a foundational data architecture in-house and the strategic contribution that has to the data maturity journey of the council is just as valuable as the technology that is actually deployed.

7.1 Platform Components for Initial MVP

Data Acquisition Layer: Moving data from disparate sources, including batch, real-time streaming, and event-driven systems as well as support for API based data integration covering Data-as-a-service capabilities. Responsible for loading the data from the data sources onto and into other data platform components, checking the data quality and storing the data in landing or staging areas of the storage layer. Covering both ETL (Extract-Transform-Load) and ELT (Extract-Load-Transform) scenarios where required.

Processing and Transforming Layer: To ensure well-formed, high-quality, and complete data. Pipeline capability to transform data from raw through to enriched and then curated and stored in the correct data model. Along with SQL based tools in this layer, no-SQL languages such as R and Python are commonly used for complex manipulation of data too. Processing includes data validation and cleansing, normalisation, de-normalisation, transformations such as structuring using joins and unions, augmentation and applying business logic to the data.

Storage Layer - Data Lakehouse: A unified storage and compute capability for data built upon data-lake storage. Supporting a wide range of formats, structures, and data types at lowest possible cost while meeting security and governance requirements. It combines the low-cost storage of a data-lake with data management features such as ACID (atomicity, consistency, isolation and durability) transactions and compute capability normally found in data warehouses, thus relieving the need to deploy a fully-fledged data warehouse. As well as supporting BI workloads

it is well placed to cultivate Machine Learning workloads. A lakehouse can incorporate much of the capabilities of the processing and transforming layer and can support the Mesh paradigm holding data 'products' from different domains separately from each other. Commonly data in the lakehouse is also placed in different folders or layers according to the degree of refinement such as the following:

Bronze (Raw Data): landing area for data without any transformation; only for performing Extract-Load (EL) operations. This layer supports any type of data.

Silver (Query Ready/ Enriched): Similar to staging databases, for performing some transformations to get data cleansed and standardised. Possible scenarios such as deduplication; also this is where data coming from different sources is merged.

Gold (Report Ready/ Curated): This is the data warehouse aspect with its dimensions and fact tables. This layer is for performing Extract-Transform-Load (ETL) operations.

Modelling, Analytics and Presentation Layer: These are the engines to query data models and serve analytics to the users, presenting visual interactive dashboards and more non-visual forms of data outputs which can kick off new cycles of data discovery, curation, modelling, and consumption. In our case these functions will be primarily implemented by exploiting Power BI Premium, in the form of datasets and dashboards for all areas of council business. The Power BI platform should enable staff at operational and strategic levels to intuitively interrogate and explore data in a self-service manner, empowering them to make data driven decisions themselves and as mentioned much progress has already been made in rolling this out.

In more limited scenarios where further complex modelling is required especially in areas such as health Informatics, other and services are required over and above the modelling capabilities found inherently within Power BI functionality and SQL. Analysts who are currently using such no-SQL tools (e.g. R and Python), locally installed on laptops and desktops, need to be provided cloud-based alternatives so that they are not so restricted by hardware and infrastructure when it comes to processing large datasets locally.

The full data production-line must be modernised which includes the right tools for informatics and business intelligence production at the analyst end of the data production-line, as often platform infrastructure can take priority over client-end modernisation. Along with such modernisation of analyst tools, there needs to be further adoption of best practices and adaptation of analyst workflow patterns and processes to suit such technology. Use of source/version control Analytics and Data Science tools, continuous deployment, test driven development and automation etc are paramount to fully exploiting the technologies while ensuring they remain supportable and sustainable without creating technical debt over time. Furthermore, this layer needs to support API's that can be used to run interactive reports on data.

7.2 Platform Components Beyond MVP

Moving on from a successful MVP for the Supporting Families programme, the following further capabilities will need to be enhanced or implemented to progress further on the organisations data journey.

Data Catalogue: Repository to capture technical, business and administrative metadata of all data products and information assets within the organisation. It helps users discover what data exists, understand what the data means, describes data lineage and relationships with other datasets, and defines who owns the data. It can act as a tool for managing information governance and security of assets (classification/confidentiality/redaction) as well as data lifecycle management enabling greater automation of data retention, archiving and deletion rules. We will use standard vocabulary[5] to represent metadata here.

Data Dictionary: This defines domain-driven design entities and their relationships to one another at a high level. (For example, how a 'person' relates to a 'household', 'property', etc.)

[5] https://www.w3.org/TR/vocab-dcat-2/

Master Data Management: MDM within a mesh architecture defines where the best version of the truth is for important domain entities across the organisation and ensures that identifiers and other key data elements about those domain entities are accurate and consistent organisation-wide. Fuzzy matching algorithms and Machine Learning algorithms can be used to support the linking of data entities helping to embrace and manage duplicated data rather than trying to eliminate it. It attempts to describe how critical data entities flow through various applications and processes across the organisation e.g. A Citizen Index for matching people across line of business systems.

Geographic Information Systems (GIS): The tools required to augment, analyse and visualise data which refers to any kind of geographic location i.e. geospatial data. Geographic data and information is central and key to all operations within local government; there is nothing the council does that does not have a spatial context. This is because a council's core and fundamental purpose, is the governance of a defined geographic area.

There is a need to review what GIS tools are used across the council and ensure that there is rationalisation as well as more standardisation to develop common patterns to integrate GIS capability into all data initiatives (where location is an important characteristic); driven by interoperability mechanisms with wider platform components e.g. mapping tools more integrated with Power BI. Furthermore, aside from the technology a wider Geospatial roadmap will be developed to advocate and formalise the best use of location data across the council.
The council also needs to be sympathetic to, and adopt national geographic data standards and strategies, such as the UKs Geospatial Strategy[6], written by the Geospatial Commission.

Data Virtualisation: Executing distributed queries against disparate data sources that are virtually integrated as a sematic layer (sometimes known as a Logical Data Fabric) which can talk to a wide range of data platforms and allow connections from any data consumers. This requires adapters to data sources, a metadata repository and a distributed query engine that can provide results in various formats (e.g., API, JDBC) for consumption. It allows consumers to access data through the semantic models which are decoupled from data location and physical schemas.

Artificial Intelligence/Machine Learning: Establish Artificial Intelligence capabilities by employing machine-learning techniques to develop predictive models without being explicitly programmed. Learning through sample and historical data to enhance areas such risk-stratification, automation, earlier interventions and better prediction and targeting.

The recent widespread proliferation and advancement of AI technologies such as Large Language Models (e.g. ChatGPT) and other generative AI algorithms, have demonstrated the power and increasing range of applications for AI technologies. Guidelines, policies and procedures for the ethical and responsible use of Artificial Intelligence for council business, will be drawn up and an AI Strategy will be formulated taking into consideration the potential social, ethical, and legal implications of this technology.

8 Working in Partnership

Although local authority data can be quite diverse and comprehensive it will not cover all needs and all circumstances, and this is clearly illustrated by the Supporting Families scenario already mentioned. As such, working with other organisations to deliver joint datasets, infrastructure and agreed outcomes is a key element of the success of data led decision making. The same 'product thinking' and other aspects of the Data Mesh Paradigm will apply to collaborative and co-created data 'products' just as they do to internal initiatives as far as possible.

8.1 Partnership structures

The citizens residing in Cambridgeshire and Peterborough are no doubt consuming public services from a multitude of organisations and agencies and so supporting the right design, development and commissioning of those

[6] https://www.gov.uk/government/publications/unlocking-the-power-of-locationthe-uks-geospatial-strategy

services means working together with these agencies to combine data (legally and safely) to get the full picture. The council already works with other agencies to share data where it is justified and appropriately governed and also works together on joint analytics projects and these important formal partnership structures will continue to be key to driving forward with more holistic insights which are of benefit to the lives of people living in the county and beyonf.

Important formal partnerships that are focused on analytics and insight already include:

- Sustainability and Transformation Partnership (STP) Digital Enabling Group
- STP Health Analytics Community and ICS Intelligence Function and Population Health Management
- Regional performance partnerships in adults and children's services
- Cambridgeshire Insight Steering Group

This strategy proposes using traditional structures such as the Joint Strategic Needs Assessment (JSNA) to support new developments of population level information on a wide range of issues. This will involve working closely with data and informatics teams in health, housing, police and other councils etc. The approach will be focused on people, places and systems to engage and align delivery to support better outcomes driven by joint data initiatives. Service governance and delivery partnerships such as the Health and Wellbeing Board, Community Safety Partnerships, the Greater Cambridge Partnership, Delayed Transfers of Care Programme Board etc, will have strong coordination roles to enable this.

The Public Health Intelligence team within the joint Public Health Directorate (CCC/PCC) and the Policy and Insight team will be particularly key in supporting partnerships. Developing new and existing relationships, they have been working for some time with local systemwide colleagues to extend data sharing across organisations, providing support for population-level planning and commissioning of services.

8.2 Joint data initiatives

Many of our most complicated challenges are about sharing data across organisational or system boundaries, e.g. child birth data, health and dental checks, health / social care interface services such as Occupational Therapy and Assistive Technology.

The technical resources of the council as large-scale providers and commissioners of public services is significant. This arises from the breadth of services that are provided by local authorities and our key role in commissioning public health programmes. The Data Mesh architecture and capabilities described in this strategy, will provide more interoperability for council data infrastructure and assets to be used across partnerships and open more opportunities for collaborative initiatives.

The council will aim to share information in the following ways in particular:

Work with partners to develop truly shared records
These will need to be focussed on particular needs and will be most useful for co-ordinating micro decisions made by operational managers e.g. ICS Shared Care Record (Local Health and Care Record) - primarily to support coordination of health and care provision for individuals.

Provide partners with population level datasets
Specific population wide information but not usually held at person or record level, used for the evaluation of an idea or strategy through modelling or hypothesis testing e.g. something akin to the Kent Integrated Dataset - primarily to support population level health and care service planning.

The latter approach is a key dependency for Communities programmes and will be needed to drive this forward in local areas.

8.3 Integrated Care System and NHS Data Platform (DSCRO)

The Cambridge and Peterborough Integrated Care Board (ICB) has a requirement (from NHS England) to have a population health and planning data platform, with business intelligence tools, by April 2024. They are working towards procurement of a platform and tools but in the short to medium term, population health work will be using information stored in the NHS Cambridgeshire and Peterborough DSCRO[7] data warehouse. This contains patient-level data from secondary care, mental health care, ambulance services, NHS community services, as well as information on births and deaths in our area. It will shortly also contain primary care data (from GP practices) and is also taking in the first data from Cambridgeshire and Peterborough local authorities in the form of some information on users of LA-provided adult social care. The data is held in a secure warehouse, regularly updated, and can be interrogated using SQL to extract and link patient level data from various sources.

The council is part of the Integrated Care System (ICS) which is a wider partnership bringing together NHS organisations with local authorities and other partners to plan services and improve health together across Cambridgeshire and Peterborough. As ICS partners, in January 2023 staff from Public Health Intelligence and Policy and Insight teams will be given access to this DSCRO data warehouse (with certain reasonable restrictions) which will enable more work joint data work with NHS colleagues on projects which are relevant for both the NHS and the local authorities.

The DSCRO warehouse is a secure system which can manage access appropriate to each user. In the medium term, the hope is to include more data from the local authorities in order to enrich population insight and develop better prevention services to keep residents healthy, as well as understand what healthcare services are needed and where. The potential uses for wider population insight are huge; the warehouse could also include some policing data, fire service data, and data from lower-tier local authorities, in particular housing data and benefits information, as well as low-level geographical information based on the 2021 Census and other sources.

It is not envisaged as a universal storage place for every piece of local authority data, but information leads across the system need to be aware of its existence and the potential to link health data to council data (and council data to other council data or other external data). It should allow new levels of population insight with great potential benefits for insight and understanding across both the NHS and councils (and other public sector bodies).

As well as developing a population health data management and insight capability, the local ICS are developing their own data strategy and so the council will continue to work closely with the wider health and care system to both engage and influence strategic direction and avoid real duplicated effort where possible across the public sector. This data strategy seeks to ensure that collaboration happens as seamlessly as possible and in scenarios where the DSCRO warehouse seems the best fit for particular council data 'products' and initiatives to reside, this will be evaluated and utilised exploiting existing wider system capabilities as much as possible.

8.4 Smart - Cambridgeshire

The 'Smart' work across Cambridgeshire looks to leverage new and emerging technologies and data to help address help address the challenges faced by the area, congestion, poor air quality, climate change adaptation and mitigation, public sector service delivery and improving the quality of life for residents across Our communities. The work is delivered in two parts.

Greater Cambridge Partnership – Smart Workstream

Smart Cambridge[8] supports and enables the Greater Cambridge Partnership (GCP) to take advantage of the opportunities provided by technology. The Smart team help to ensure that our investments and the legacy we leave behind are appropriately future focused by considering existing and emerging technologies.

The work of the Smart Cambridge team is focused around five key themes, namely to:

[7] https://digital.nhs.uk/services/data-services-for-commissioners/data-services-for-commissioners-regional-offices
[8] https://www.connectingcambridgeshire.co.uk/smart/

- Use data and technology to support a shift from private car usage to sustainable forms of transport.
- Use data and technology to improve the management of the highway network.
- Improve the range and quality of data available to GCP, and explore innovations to help GCP and its partners make best use of the data available
- Develop strategies and approaches to ensure that new communities take advantage of data and technology to promote sustainability
- Keep GCP abreast of external developments and potential collaborations to advance towards GCP's objectives

To support this work Smart Cambridge has been experimenting with an intelligent city data platform in collaboration with University of Cambridge. The platform consists of the following:

- LoRa Network which is a low powered communication network for sensors
- IoT sensors – Bluetooth traffic sensors, Air Quality Sensors, in-Building CO2 sensors etc
- Real-time data – Bus and Train data, Parking data
- Data platform which collects a number of streams of real-time data and allows access through an Open API as well as displaying the data through visualisations.

The platform, over the last few years, has demonstrated the value of brining such data together and has underpinned a number of projects including air pollution monitoring, enhanced publishing of real-time travel information and enabled policy makers to get basic answers to transport related questions.

The Smart Cambridge Platform has supported the Smart Workstream to use collate and use various data streams and a number of products such as analytical tools and travel planners have been built on top of it. However as this is a joint project with the University the development and maintenance is on 'best efforts' and it isn't suitable for a 'business as usual' deployment. The Smart Workstream is exploring with partners what data infrastructure will be needed to support the GCP programme particularly as it begins at scale deployment of technology. An example of this is the at scale deployment of VivaCity traffic sensors along with partners 80 sensors have now been deployed, the challenge is how this data is cleaned, structured, analysed and shared with other public sector bodies. A platform would help to automate these processes, greatly reducing the effort needed to extract value from the data.

Connecting Cambridgeshire – Smart

The Connecting Cambridgeshire Team have been funded by the Cambridgeshire and Peterborough Combined Authority (CPCA) to deliver:

- A LoRa network in parts of Huntingdonshire, East Cambs and Fenland
- Sensor pilots – flooding and Air Quality

It is likely that this work will scale up to explore other areas of sensor deployment that could help and support service delivery with the county council. Currently these deployments use the sensor supplier's data platform and this tends to mean data is siloed. Connecting Cambridgeshire have been working with GCP, CPCA and the County to look at how a data platform could support the aggregation and sharing of the data they have started to collect.

The data capabilities described as part of this strategy will be developed to incorporate the platform requirements for both parts of the Smart programme of work. The ability to manage large amounts of unstructured streaming data such as that demonstrated in the Smart Cambridge platform is of critical value to our overall data objectives, providing a solid foundation for other such IoT initiatives to be included and expanded going forward e.g. better and wider use of assistive technology, telecare and other smart home monitoring technologies resulting in enhanced preventative care provision.

9 Open Data

The council is committed to community engagement and from a business intelligence perspective, the publication of open data is an important contribution to this goal. Data collected and processed as part of the delivery of public services is a public asset, and so it should be made available to support transparent government and because it is of instrumental value to the people in our communities analysing and visualising data to support the best possible decision-making as well as fostering innovation from data. Data will continue to be published on our **Cambridgeshire Insight Open Data Portal**[9], and an **open data workplan** will be developed in partnership. In particular data domains, where Services are not dealing with sensitive data, then an 'open by default' approach will be considered for those service areas, with principles and policies formulated to enable and foster this.

Open data is usually made available in a relatively 'raw' way, for the user to access it and use as they see fit. However, as part of the open data agenda, the main website, **Cambridgeshire Insight**[10], developed by both CCC and PCC on behalf of a partnership of local public sector organisations, was setup to provide more intuitive and accessible interactive content with accompanying narrative and context.

The platform supports the visualisation of data relating to a variety of needs and services, as well as the publication of analysis in a variety of mediums, both interactive and static. For example, Cambridgeshire Insight can be used to host Joint Strategic Needs Assessment documents, interactive maps of housing information, crime or population change, as well as profiles for wards or other administrative areas which combine population, economy, Census and housing data together in an easily accessible and publicly available manner.

The continued development of Cambridgeshire Insight to support Communities programmes and other place-based public service delivery will be laid out in a **development roadmap**.

10 Unstructured Data

Most of our unstructured data resides outside of line of business systems i.e.

- Emails in both personal and shared mailboxes
- Files stored locally on PCs and laptops
- Files stored in network storage such as teams shared areas and 'home directories'. This data resides on a Storage Area Network (SAN) which is being decommissioned in August 2023.

During the adoption of Microsoft 365 all emails had been moved to the online version of Exchange as part of the migration process. A project is also now underway to update all staff to a new version of Windows managed by Microsoft InTune which includes migrating their locally stored files and personal network folders to OneDrive online.

The rest of the files held in shared network folders residing on infrastructure in the Sand Martin House data centre (i.e on the SAN) remain the most challenging aspect of this migration to Microsoft 365 storage. These files will all be migrated to Microsoft Sharepoint/OneDrive and a workable process to move this data has been piloted already and now needs to be initiated for the whole authority and completed before SAN decommissioning.

As more and more unstructured data is moved into the cloud it opens up opportunities to apply and use technologies on that data for processing and analysis, transforming it and extracting out of it metadata of a more structured nature. We hope to leverage Artificial Intelligence and Machine Learning capabilities to apply techniques such as Natural Language Processing (NLP), Text Mining and Optical Character Recognition (OCR) to not only provide insights into the unstructured data, such as sentiment analysis or visualisations, but also to support

[9] https://data.cambridgeshireinsight.org.uk
[10] https://cambridgeshireinsight.org.uk

business process transformations. For example, using Robotic Process Automation (RPA) to create intelligent document processing solutions to convert documents like invoices, contracts, agreements etc into structured formats and automate elements of processing them. Also such capabilities will allow better governance and control of unstructured data through automation, supporting aspects of data management such as retention, availability, classification, correct storage 'location', duplication, validity, currency etc. Much of this will be progressed naturally by exploiting the Microsoft 365 ecosystem together with Microsoft Power Platform capabilities, allowing and encouraging low-code and no-code development to be pervasive across the organisation and not limited to just the IT and Digital service.

Appendix A - Glossary of Terms

Application Programming Interface (API): A software intermediary that allows two systems to talk to each other and exchange data. APIs are an accessible way to extract and share data within and across organisations.

Continuous Integration/Continuous Deployment (CI/CD): CI is a modern software development practice in which incremental code changes can be made frequently and reliably. Automated build-and-test steps triggered by CI ensure that code changes being merged into the code repository are reliable. The code is then deployed quickly and seamlessly as a part of the CD process. The CI/CD pipeline refers to the automation that enables the incremental code changes from developers' machines to be delivered quickly and reliably to production environments resulting in increased early defect discovery, better productivity, and providing faster release cycles. CI/CD are DevOps practices and contrast with traditional methods where a collection of software updates were integrated into one large batch before deploying the newer version.

Data Cleansing: Reducing irregularities and omissions in data to provide credible data for all uses such as use in applications or reporting.

Dataflows: In the context of Power BI, Dataflows are basically a self-service data integration tool available in the Power BI service that can be used to fetch data from various data sources and create a data model in the cloud based on the schema of the datasets. An advantage of using the dataflows is that they can also be reused within the organisation, and as such, you can create modular ETL pipelines to prepare datasets. Dataflows can optionally be hosted on Azure Data Lake Gen2 storage for greater access and control.

Data-lake: A highly scalable storage repository that holds large volumes of raw data in its native format, without transformation, until it is required for use. Data-lake data often comes from disparate sources and can include a mix of structured data from relational databases (e.g. rows and columns), semi-structured data (e.g. CSV, logs, XML, JSON), unstructured data (e.g. emails, documents, PDFs, sensor telemetry) and binary data (e.g images, audio, video). Data is stored with a flat architecture and can be queried and transformed as needed.

DataOps: A cross-functional set of practices, processes and technologies to operationalise data management and data architecture support provision with emphasis on communication, collaboration, integration, automation, speed and quality. DataOps seeks to improve data operations in the cloud in a similar fashion to DevOps and borrows similar practices.

Data-warehouse: A data storage architecture to process and transform data for advanced querying and analytics, normally in a structured environment such as a relational database. Designed to hold data extracted from transaction systems, operational data stores and external sources where required. The warehouse then combines that data in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs.

De-normalisation: Combining data from multiple tables into a single table so that it can be queried more efficiently for reporting and analysis.

DevOps: A cross-functional set of practices, processes and technologies combining elements of development (Dev) and operations (Ops) designed to increase an organisation's ability to deliver applications and services faster than traditional software development processes. Exploiting the agility found in cloud services together with adoption of agile and lean methods of managing people, process, and technology applied to the full software production life-cycle.

DSCRO (Data Services for Commissioners Regional Office): A service provided by an external NHS organisation that allows storage and management of individual-level and identifiable data. ICSs are not allowed to hold this type of data directly themselves.

Extract Transform Load (ETL): A three-phase process to extract data from different sources, transform the data into a usable and trusted resource, and load that data into systems, often data warehouses, for further processing or analysis by end-users downstream to solve business problems.

Extract Load Transform (ELT): An alternative to ETL in which data is extracted and loaded and then transformed. This sequence allows the preloading and storage of raw data to a place where it can be modified later. ELT is more typical for consolidating data in a data-lake, as cloud-based data-lake solutions are capable of scalable processing and transformation on demand.

Minimal Viable Product (MVP): A basic, launchable version of a product that supports minimal yet must-have features. It is built with the intent to enable faster time to adoption, obtain early feedback, and achieve 'product-market' fit from early on. The MVP concept is perceived as a combination of the 'minimum essentials', something that has the basic features to satisfy the initial users but which can then be built upon incrementally as feedback is received.

Natural Language Processing (NLP): Uses artificial intelligence to give computers the ability to interpret, manipulate, and comprehend human language such as voice and text, e.g. emails, text messages, social media newsfeeds, video, audio. Intent or sentiment in the messages can be analysed and even the provides the ability to respond, in real-time to messages, as part of conversation.

Normalisation: Process of reducing redundancy and inconsistency in data as well as cleaning the datasets of unused data.

Optical Character Recognition (OCR): The process that converts an image of text into a machine-readable text format that can then be processed as structured data. Often uses a combination of machine learning and computer vision algorithms.

Robotic Process Automation (RPA): Software to automate tasks and processes within businesses via scripts that emulate human interaction with application user interfaces. Often the scripts are produced using low-code and no-code platforms which are designed to be used by non-technical users and technical users alike.

Storage Area Network (SAN): A high speed network of storage devices that can be accessed by multiple servers or computers, providing a shared pool of storage space typically covering all enterprise storage requirements, from network file shares for users, to storage for servers.

Text Mining: Identifies facts, relationships and assertions, that would otherwise remain buried in the mass of textual data; by transforming (unstructured) text in documents and databases into normalised and structured data suitable for analysis. Results are often presented using HTML tables, mind maps, charts, etc. Text mining uses a variety of techniques to process the text, one of the most important of these being Natural Language Processing (NLP).

Version Control: In the context of software engineering, version control, also known as source control, is the practice of tracking and managing changes to software code. Version control systems (VCS) are software tools that help development teams manage changes to source code over time and support collaboration on software projects. A VCS keeps track of every modification to the code in a special kind of database and can often manage different iterations (with different features) of the same software being developed simultaneously, as different 'branches', which are eventually amalgamated. If a mistake is made, developers can turn back the clock and compare earlier versions of the code to help fix the mistake while minimising disruption to all team members.

Record of Change

Version no.	Date	Author/Owner	Description of Change
1.0	21/06/2023	Abdul Hadi	First release

Approvals

Signature	Print Name/Board	Date	Title/Role
Alan Shields			
Sam Smith			