



The Evolution of the Pure Community Module

How lessons learned from national, regional, and subject-focused use cases have been used to support inter-institutional collaboration

15th International Conference on Current Research Information Systems (CRIS2022), Dubrovnik, Croatia, May 2022

James Toon, Product Manager - Pure

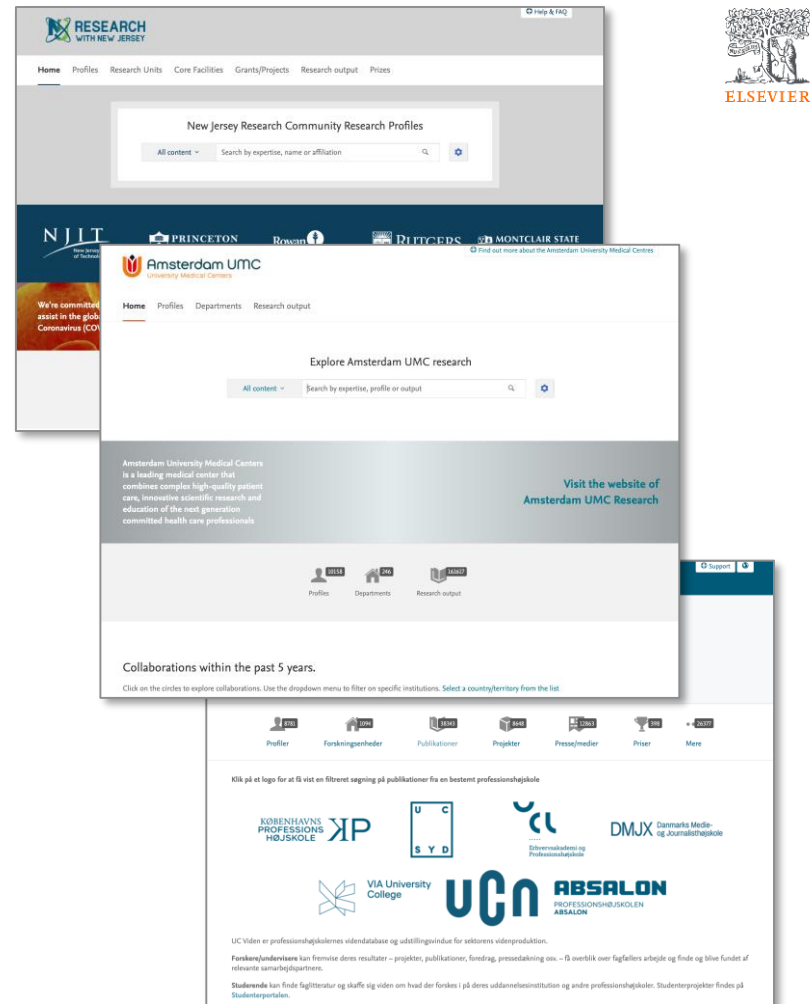


Agenda

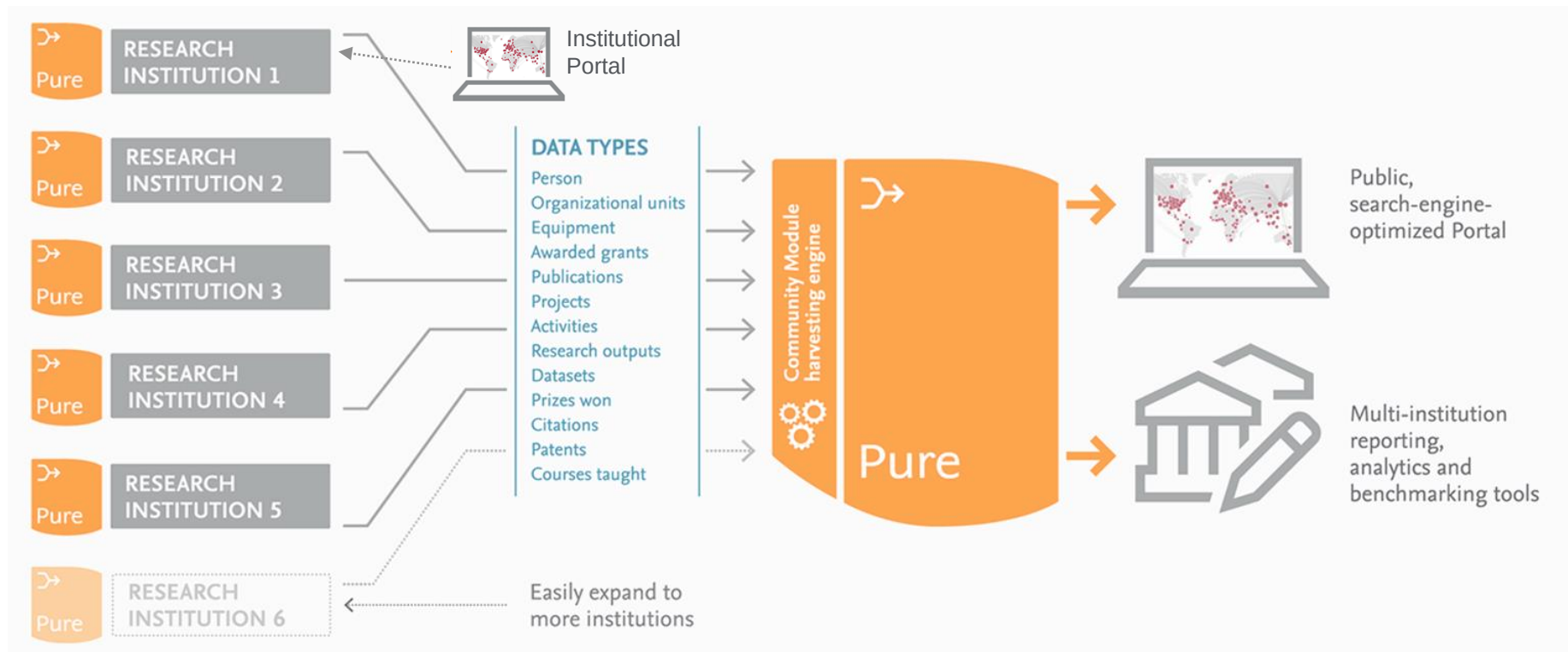
- Background
- Pure Community Module
- Evolution of the Module
- Lessons learned (summary)
- Future directions
- Conclusions
- Q & A

Background

- The Pure Community module is a service offering that provides a technological framework to satisfy the needs of inter-institutional collaboration
- The Pure community module and our knowledge of its use has developed substantially since our last presentation at EuroCRIS 2018
- Lessons we have learned through working with communities over the years has informed future direction



The Pure Community Module



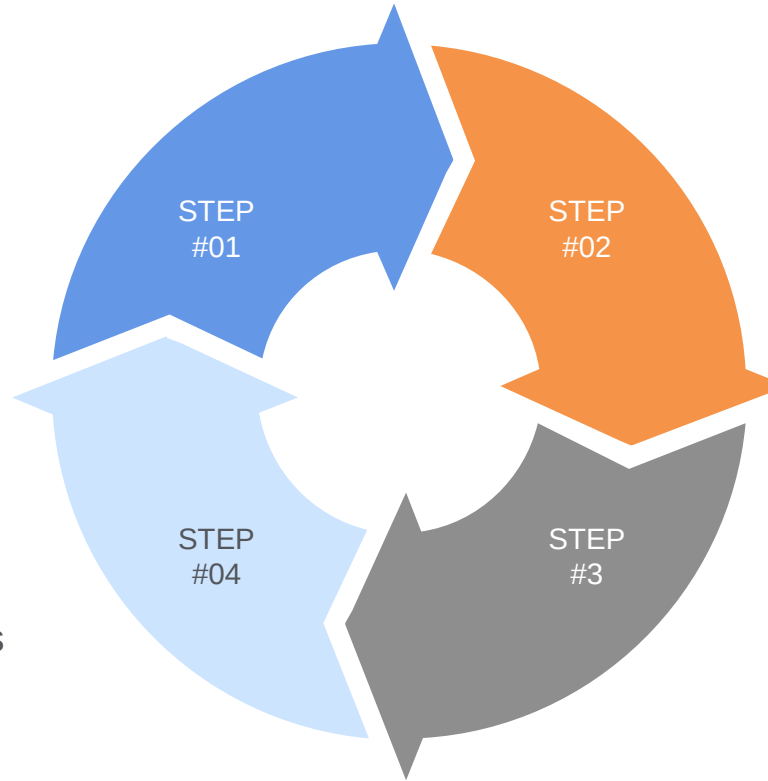
The Pure Community Process

1. Pure Clients

Content harvested asynchronously following creation or update (based on changes to object hash values).

4. Deduplication

performs deduplication tasks to reconcile data from across instances based on defined identification, validation and processing criteria



2. Harvesting and Processing

Harvester service retrieves data and data delivery service prepares changes based on defined criteria.

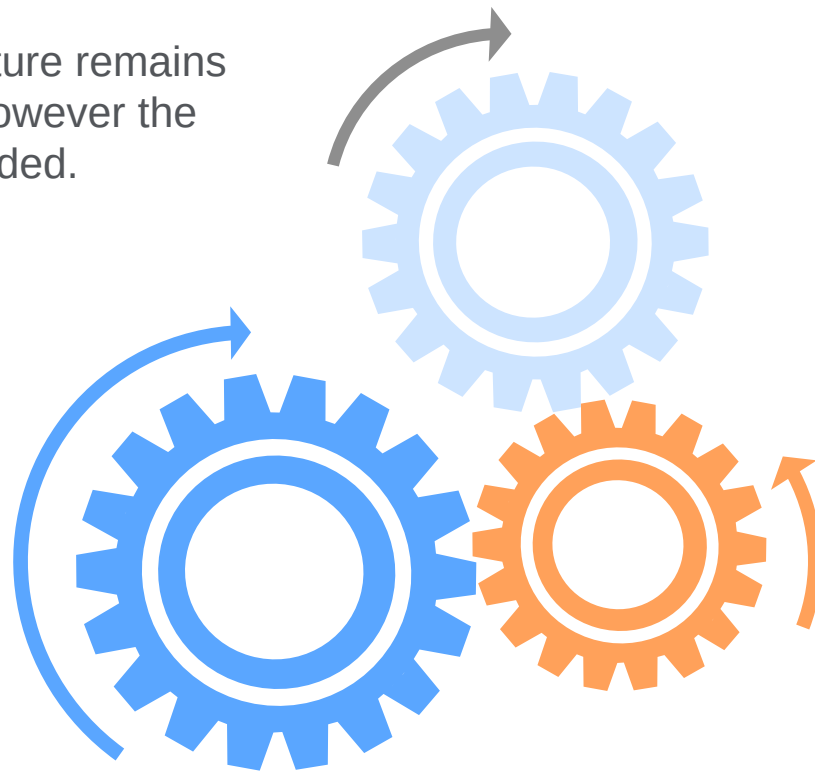
3. Synchronisation and Cleanup

Retrieves data from harvester and performs community synchronisation, covering deletion actions and data create/update operations

The Pure Community

At its core, the community module architecture remains must as it was when reported previously, however the components have been substantially upgraded.

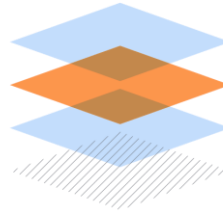
- Updates to web service in line with Pure system developments & Pure in general
- Extended content scope (type coverage)
- Significant updates to deduplication logic
- Updates to harvester architecture and data providers within the sync framework
- Messaging/event logging service to improve transparency in process



Expansion of Community Use Cases/Drivers



Multiple institutions combined with an aggregating, shared central Community instance, allowing comparative evaluation and showcasing of community expertise via dedicated portal



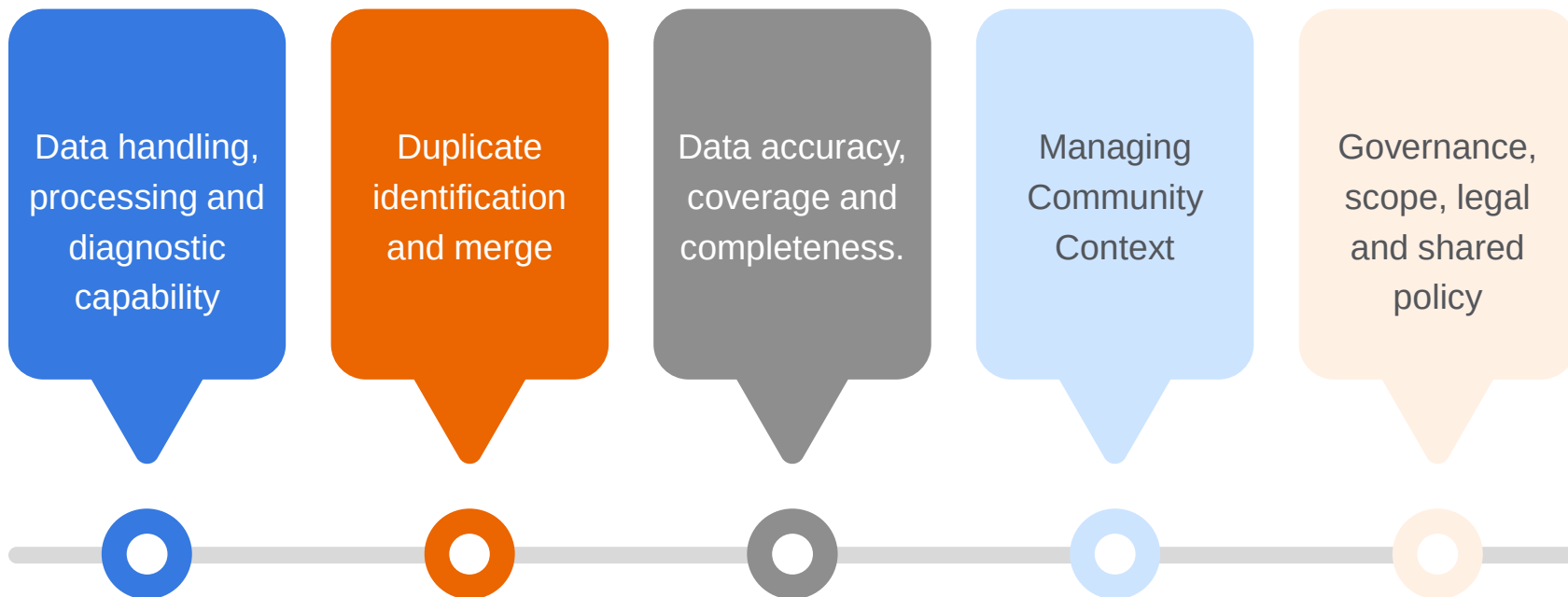
National performance management via a transparent, inclusive digital workflow, or in support of a transparent, inclusive national evaluation via data collection



Regional/State infrastructure to develop economic & collaboration impact, demonstrate expertise to foster collaboration & economic development, and Identify expertise within a region

Summary of lessons

Our review of community operations gathered evidence from development teams, working in partnership with customers and through technical and account support and led to a number of key areas of focus



Summary of lessons

Our review of community operations gathered evidence from development teams, working in partnership with customers and through technical and account support and led to a number of key areas of focus

Data handling,
processing and
diagnostic
capability

Data volumes in an aggregate community system are considerable compared to individual institutional systems. This has a direct impact on the performance and sensitivity of processes required to maintain the transfer of data between local and community instance.

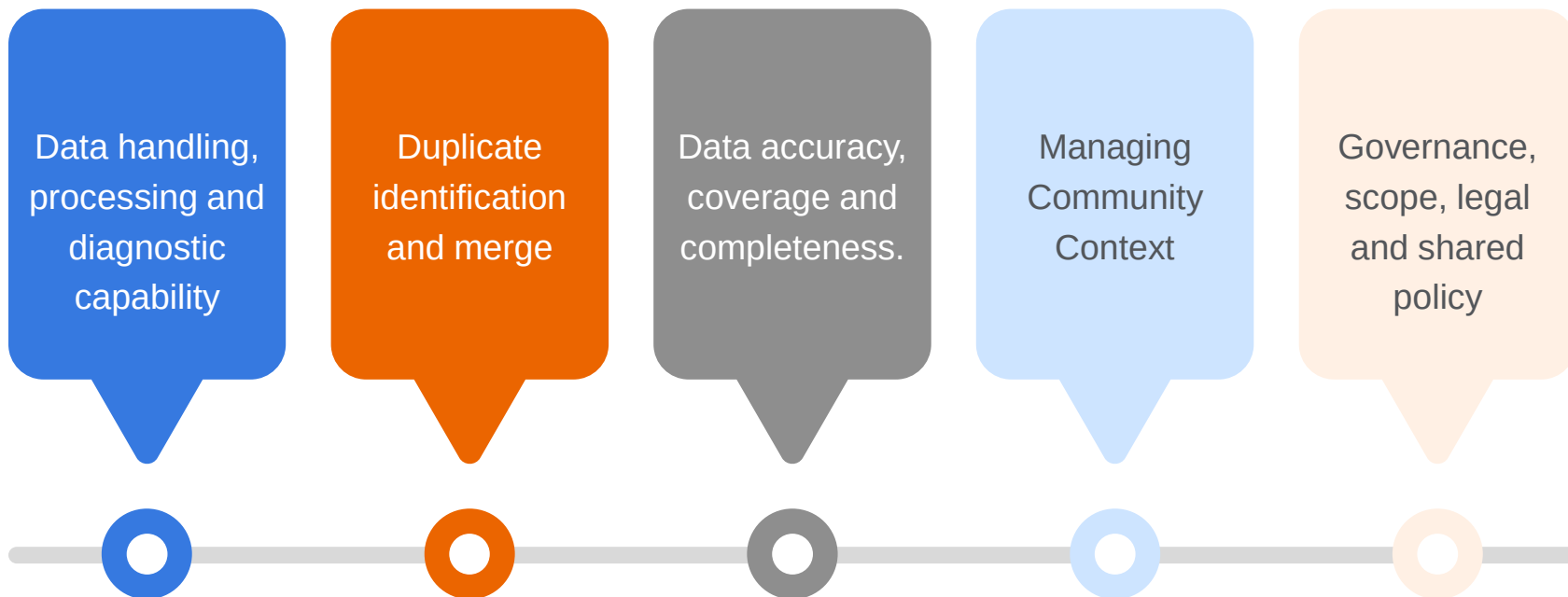
Typical publication volume for community may be >0.5M publication records or more. ~1M entities of more across all data types

Single 'run' for community example comprises 25 individual sync processes. Range of logging from 15000 (incremental) to >1M (full run) entries

Need to know where everything is all of the time – key trust component

Summary of lessons

Our review of community operations gathered evidence from development teams, working in partnership with customers and through technical and account support and led to a number of key areas of focus



Summary of lessons

Our review of community operations gathered evidence from development teams, working in partnership with customers and through technical and account support and led to a number of key areas of focus

Data handling,
processing and
diagnostic
capability

Duplicate
identification
and merge

The aggregation of multiple content sources creates a challenge in the identification, validation and merging processes across all content. A parameter driven approach to deduplication is effective however the process needs to be aligned to local policies.

Use of PIDs are hugely important

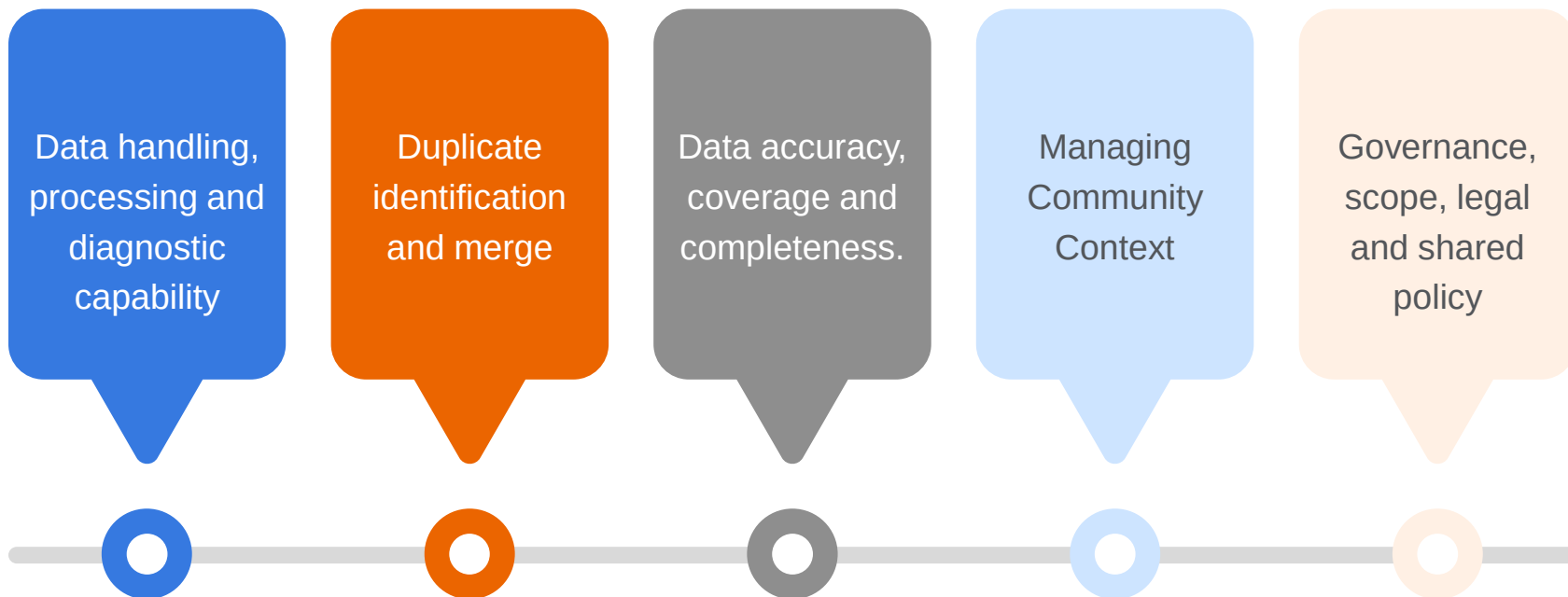
Some changes can be considered dilution, not enrichment

Merging of data with large numbers of dependencies (for example related to hyper-authorship papers)

Example 2

Summary of lessons

Our review of community operations gathered evidence from development teams, working in partnership with customers and through technical and account support and led to a number of key areas of focus



Summary of lessons

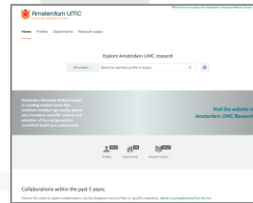
Our review of community operations gathered evidence from development teams, working in partnership with customers and through technical and account support and led to a number of key areas of focus

Community organisational structures may not be a direct representation of the combined organisational structures of the participating instances. This requires a means to make available alternative structures and/or a subset of the overall community dataset for purposes of showcasing

Amsterdam University Medical Centres

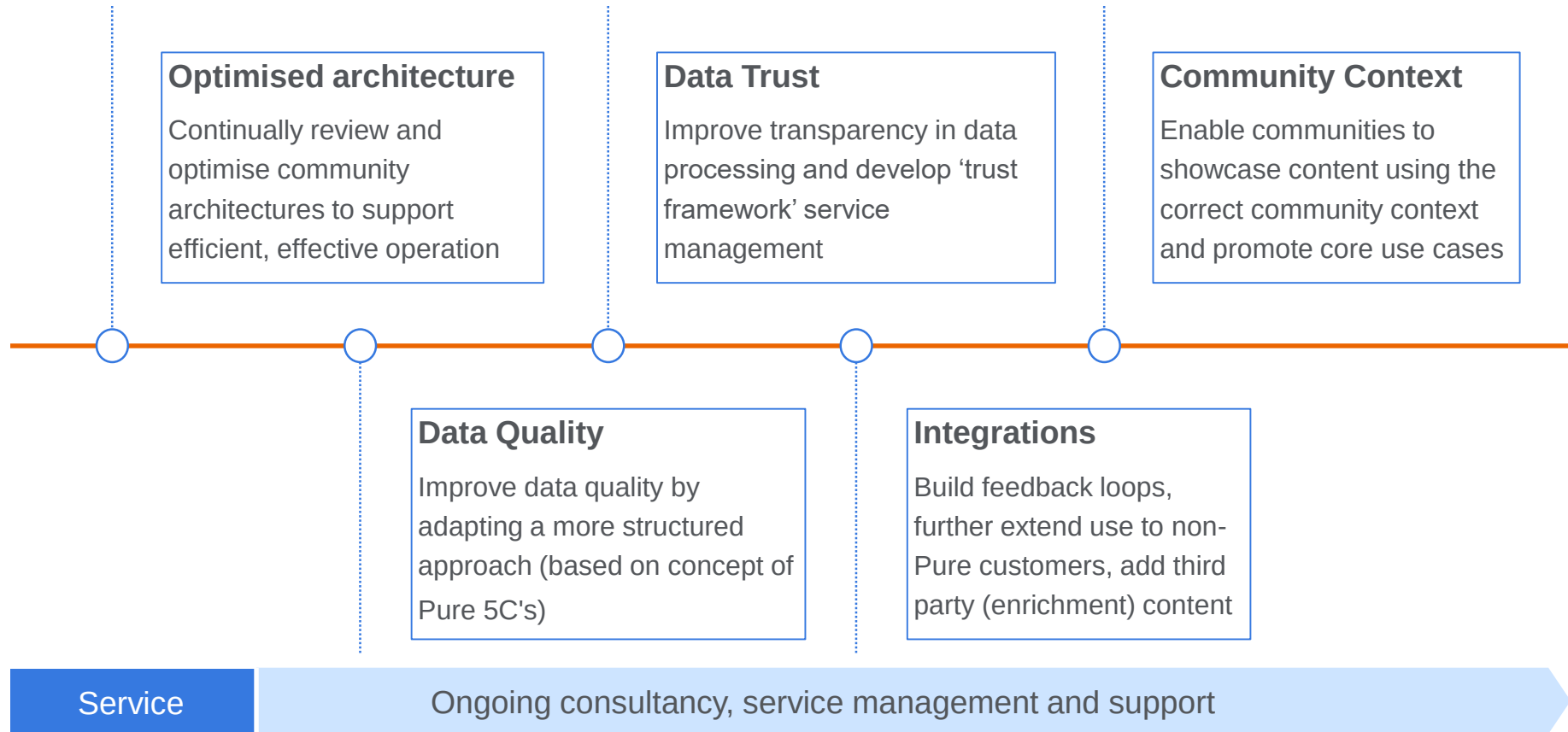
Managing
Community
Context

Governance,
scope, legal
and shared
policy

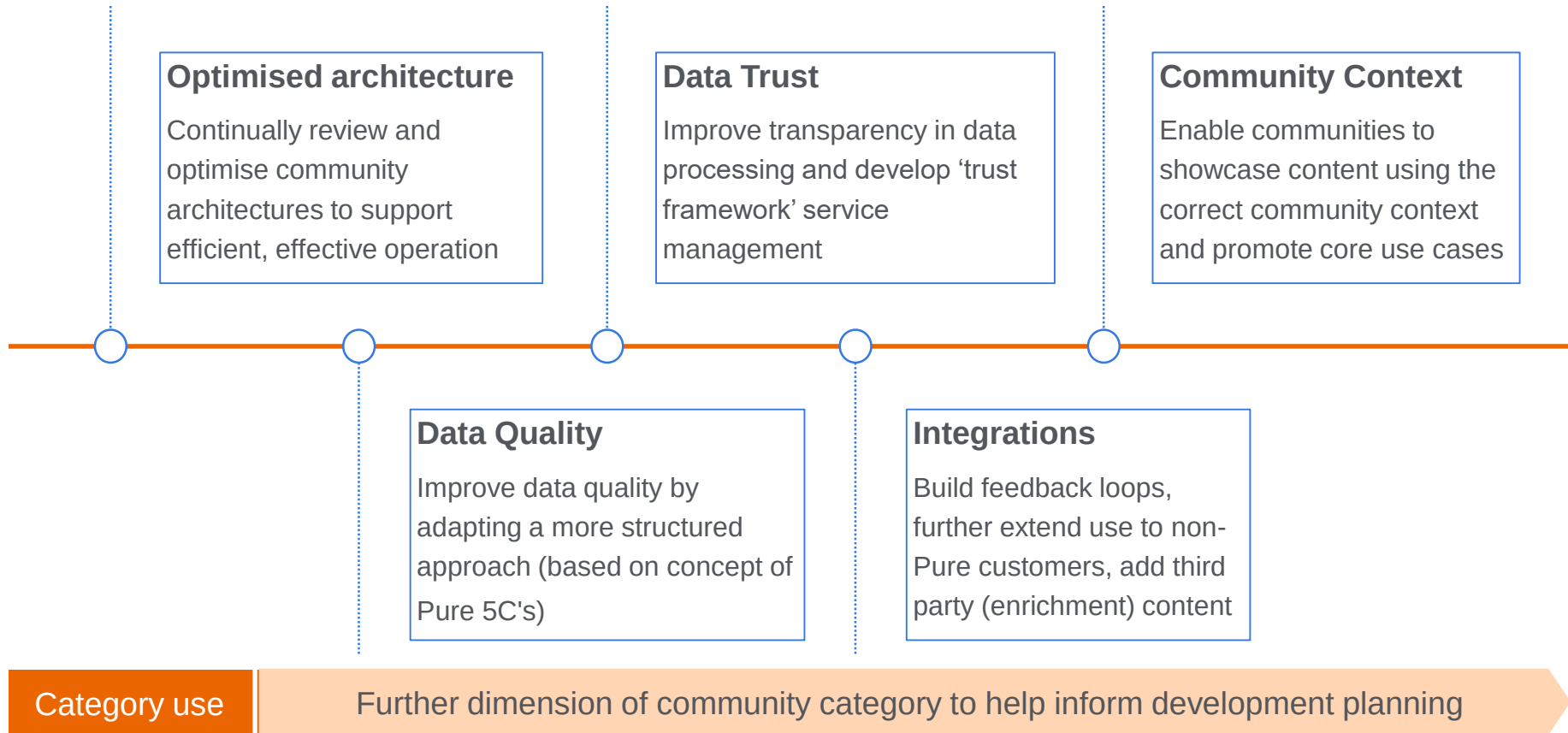


Example 3

Looking ahead - Using Lessons to Inform Future Direction



Looking ahead - Using Lessons to Inform Future Direction



Conclusions



- Community solution are established as an approach to the inter-institutional use case of CRIS/RIM systems on national and regional basis.
- We have identified several core usage scenarios and this will likely expand further. We must always learn and adapt.
- The architectures and technology supporting communities will continue to mature – new service options will start to be developed. Organisational challenges will develop new areas for investigation
- The management of community based aggregations requires a different skillset to institutional CRIS
- Community providers will continue to focus on the changing needs of community as a co-management/partnership arrangement

Any questions?

James Toon

May 2022

