SDMX GC

Using data description to automate validation with VTL

Thomas Dubois
Franck Cotton

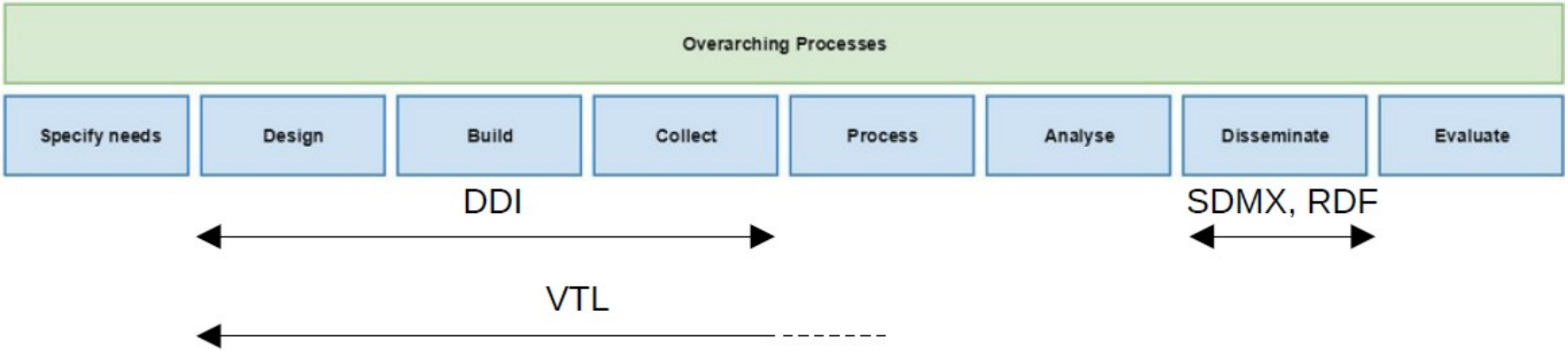
Insee

- Metadata strategy
  - Align on international standards
  - Active metadata all along the statistical process
  - Embrace open data and open source

Insee context

- Metadata standards
  - DDI for questionnaires and variable-level documentation
  - SDMX for dissemination
    - Dissemination of time series by API
    - Automation of data dissemination
  - VTL for data validation
    - Data controls and flow logic for electronic surveys
    - Reconcile multimode household survey data
    - Validate administrative data

- Metadata standards and GSBPM

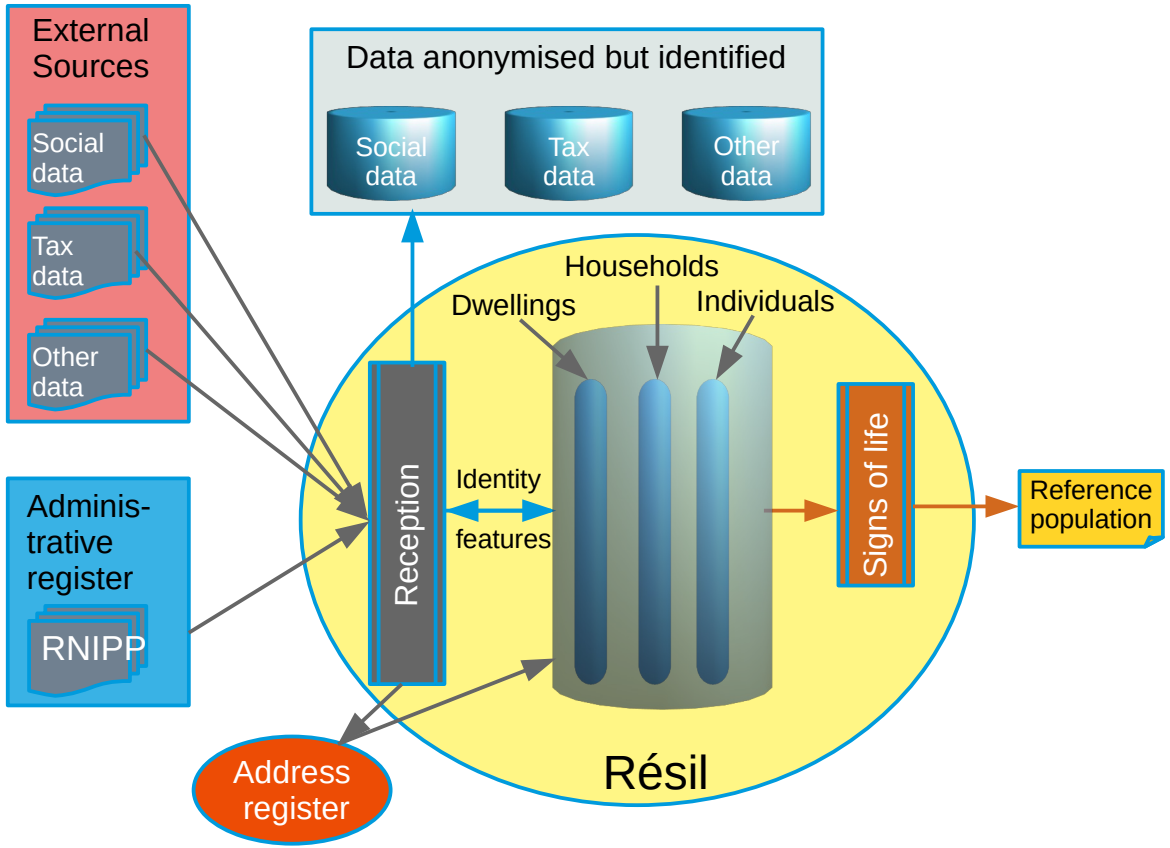


Generic use of standards in the statistical process

- Zoom on VTL
  - Validation and Transformation Language
  - Published by the SDMX initiative
  - Desirable features
    - Business-oriented, independant of technology
    - Formal grammar -> automatable

Use case

- The Résil system
  - Build a statistical register of individuals and dwellings based on the linkage of various administrative data
  - See also ISI presentation



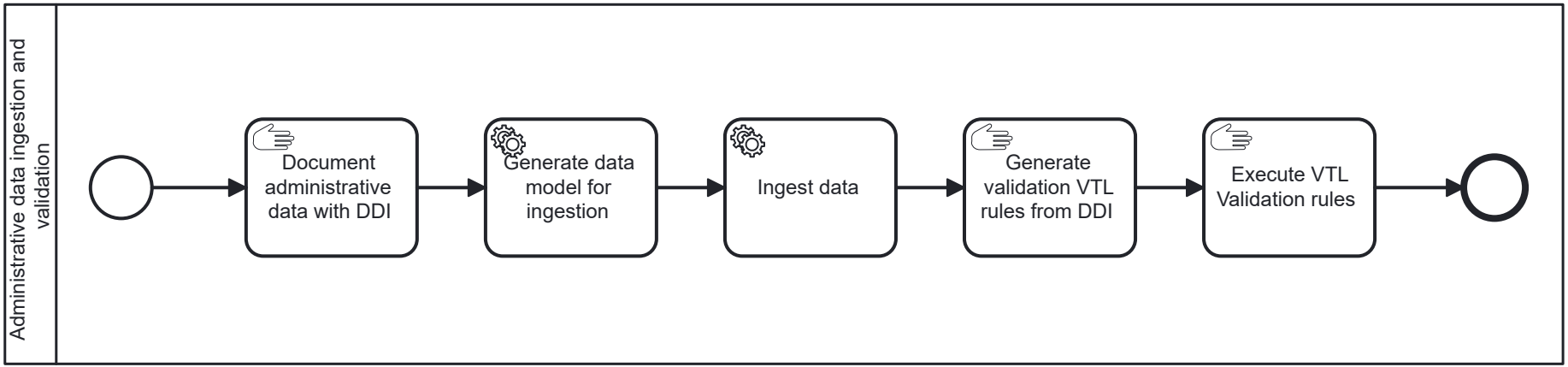
Use case

- Main objectives
  - Automation of administrative data ingestion and validation
  - Documentation of:
    - Data ingested
    - Validation rules

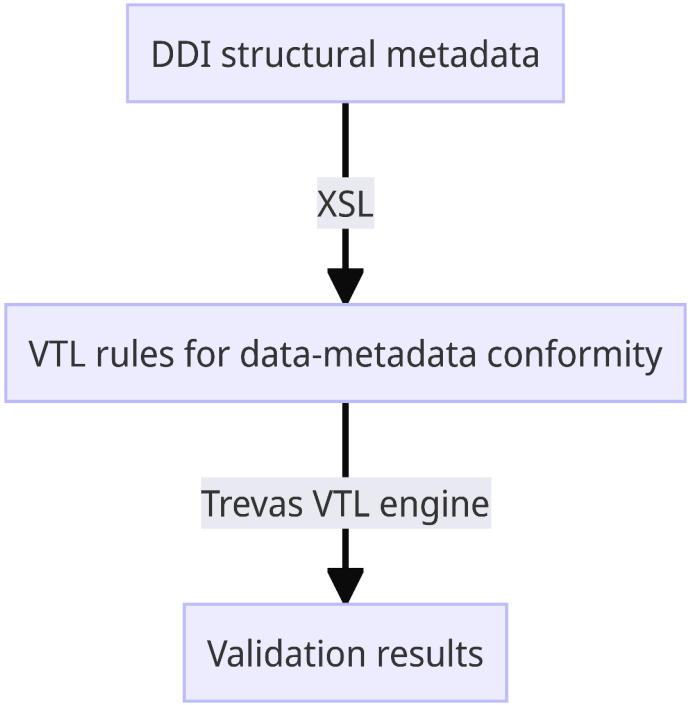
Use case

- Workflow
  - Documentation of administrative data with DDI
  - Generation of data model for data ingestion
  - Ingestion of data
  - Generation of VTL validation rules from DDI
  - Execution of VTL validation rules

- Workflow


Administrative data ingestion and validation
Document administrative data with DDI
Generate data model for ingestion
Ingest data
Generate validation VTL rules from DDI
Execute VTL Validation rules

- Technical workflow
  - DDI structural metadata
    - Entered in Colectica system
    - Exported in XML
  - Java-driven XSL transformation
    - Produces VTL **`ruleset`** objects
  - Trevas VTL engine
    - Runs validation script
    - Provides datasets of results



- Examples

```
define datapoint ruleset dpr_ETAB (variable code_decl, id_mad_etab, siren) is
    rule_code_decl : code_decl in {"11","14"}
        errorcode "Invalid code value";
    rule_id_mad_etab : between(cast(id_mad_etab, number), 1, 999)
        errorcode "Value out of bounds";
    rule_siren : between(length(siren), 9, 9) and match_characters(siren, "[0-9]*[1-9][0-9]*")
        errorcode "Invalid SIREN"
end datapoint ruleset;
```

Conclusion

- Proof of concept conclusive
  - Activation of structural metadata
  - Seamless insertion in statistical process
  - Value added in terms of quality
    - Better documentation of data and treatments
    - Coherence, traceability, adaptability

Conclusion

- Thank you
- Any questions?