Assignment for Senior Software Developer Role – 4TU.ResearchData Repository

Context:
The 4TU.ResearchData repository allows users to browse and search datasets by institutions. Currently:

- Institution information is retrieved from the depositing author's affiliation (identified via login).
- If the institution cannot be determined from login credentials, the depositing author selects it from a list in the metadata form.
- Based on institutional grouping, we provide statistics (e.g., number of publications).

The Challenge:
Our stakeholders (data stewards, faculty deans) now require **statistics per faculty**, not just per institution.

- Faculty information can be derived from the metadata field "Organizations", but this field is free text and therefore unreliable.
- Additionally, many publications have multiple authors from different universities, yet statistics are currently grouped only by the depositing author's institution.
- We do collect ORCID IDs for authors, but it is not a mandatory field, so for some entries it is missing.

Summary of the user motivation:

As a data steward of 4TU*, I want to check statistics per faculty, not only at the organizational level (in the case of TU Delft institution, e.g. Faculty of Technology, Policy and Management), so that I can analyze the dataset statistics specifically relevant to my faculty.

4TU: Delft University of Technology , Eindhoven University of Technology, University of Twente And Wageningen University and Research*

Your Task:
Design a new feature that addresses these issues and improves the accuracy and usability of faculty-level statistics. Your solution should:

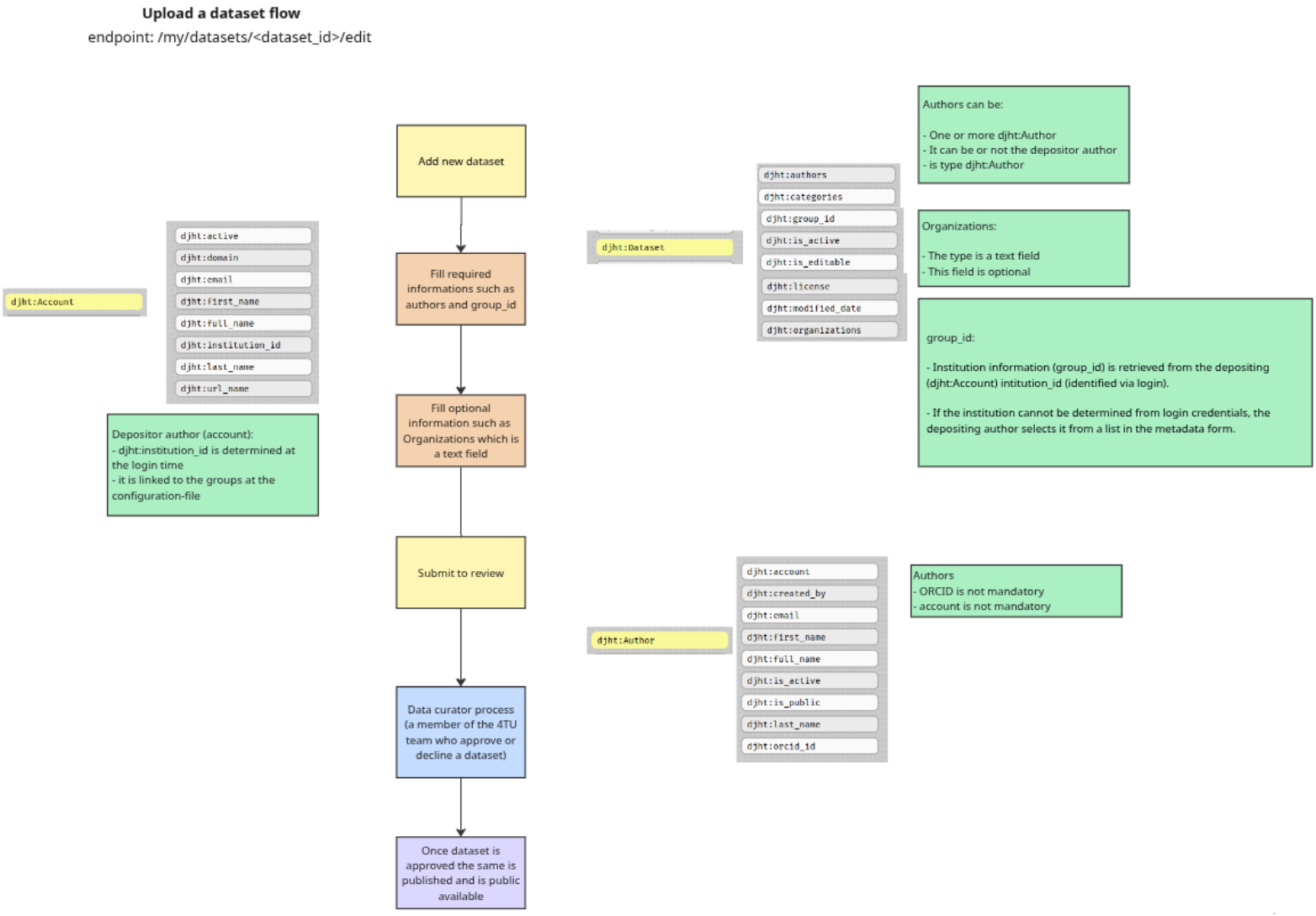
1. **Conceptualize the approach:** Explain the problem, your proposed solution, and why it is effective.
2. **Address existing data:** Suggest strategies for handling already published records with unreliable or incomplete metadata.
3. **Technical implementation:** Outline how you would implement your solution in the repository's source code (a full implementation is not required, but we provided a development environment so you can try out to check feasibility).

4. **Consider edge cases:** Multiple authors, missing ORCID, inconsistent metadata, etc.
5. **Highlight advantages:** Explain how your solution benefits stakeholders and improves the repository.

Deliverable:

Prepare a **10–15 minute presentation** that covers:

- The conceptual design of your solution
- The technical approach (you can include data model changes, algorithms, or integration strategies)
- Handling of existing data and edge cases
- Advantages and potential limitations
- When working on the task, what were your impressions of main strengths and weaknesses of the system, and how would you suggest addressing a specific weakness?



Useful links:

Dataset samples

https://data.4tu.nl/datasets/ac081516-5311-481e-b7a2-07c5ea9b1991
https://data.4tu.nl/datasets/31f3537b-4137-4ac4-bde2-e0811105921c
https://data.4tu.nl/datasets/40731af1-dfcc-46eb-93c2-b4d715676ea6
https://data.4tu.nl/datasets/49513019-dc6d-4f6e-9bc5-36bc4d8db1a7
https://data.4tu.nl/datasets/342efadc-66f8-4e9b-9d27-da7b28b849d2/5

ORCID sample:
https://orcid.org/0000-0003-1718-3109

Online sandbox:
https://next.data.4tu.nl/