Duplicate_Detection

Important links

Resource	Link	Details
Jira Epic	DBT-60646: Duplicate Detection DONE	
Meeting Notes	Weekly Meeting Notes	
Launch Date	29.07.2025	
Help Center	https://drooms.zohodesk.eu/portal/en/kb/articles/searching-and-filtering-data-rooms	

Project team (contact people)

- PM Lead: @Johannes Graf
- BE Lead: X
- FE Lead: X
- KMS Lead: X
- QA Lead: X
- Design Lead: X
- CS Lead: X

Summary

The **Duplicate Detection** feature introduces a dedicated filter in the Drooms dataroom, enabling users to quickly identify and manage duplicate or near-duplicate documents. This helps prevent misunderstandings, eliminates versioning conflicts, and streamlines content review.

Duplicates can now be identified both by exact matches (hashcode) and by content similarity. This means that the system is able to detect, for example, different scans of the same document or contract versions with and without signatures.

All duplicate and near-duplicate results are ranked by similarity, making it easier for users to review and decide which documents to keep.

Additionally, users can now highlight differences directly in the document viewer, allowing for fast comparison without leaving the dataroom.

The duplicate filter can be combined with other filters (e.g., document type, date, or redaction status), providing powerful customization when reviewing content.

User Research

During discussions with key customers,, who signed a framework contract for asset management projects, the demand for duplicate detection was repeatedly raised.
Large-scale asset management processes require fast and reliable ways to detect redundant documents to ensure accuracy in due diligence, audits, and document-heavy reviews.

Key findings

- Users want to **identify duplicates across the entire dataroom**.
- Users need to **combine this filter with existing filters** to narrow down results (e.g., “show only duplicate PDFs uploaded in the last month”).
- Users want to **delete duplicates directly from the filtered view** for better document management.

Problem Statement

Currently, users have no way to systematically identify duplicate documents in the dataroom. This can lead to:

- **Confusion**: Multiple copies of the same documents may create uncertainty about which one to use.
- **Review inefficiencies**: Time is wasted reviewing redundant documents.
- **Versioning conflicts**: Incorrect versions may be referenced during audits or due diligence processes.

Goal: Provide a simple yet powerful way to filter and manage duplicate documents, while respecting users’ access rights

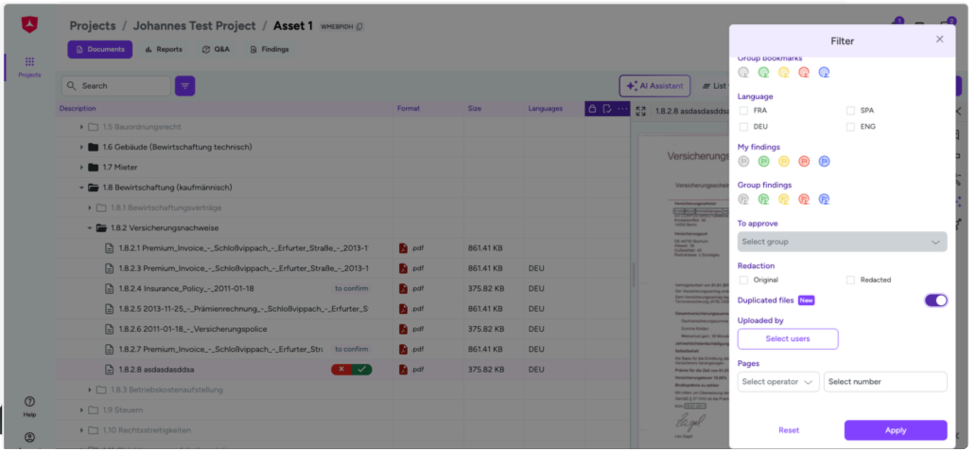
Features

Core Functionalities:

- **"Duplicates" Filter**
  Filter for duplicate and near-duplicate documents across the dataroom.
- **Similarity Ranking**
  Duplicates are sorted by similarity, with the closest matches displayed first.
- **Access-Controlled Display**
  Users only see duplicates they have rights to access.
  - Example: If there are 3 duplicates and the user has access to 2, only those 2 will be displayed.
- **Filter Combinations**
  The duplicate filter can be combined with other filters (e.g., document type, date, redacted status).
- **Clustered Display**
  Duplicate and near-duplicate documents are grouped together in the index for clarity.
- **Bulk Management**
  Users can delete one or multiple duplicates directly from the filtered view.
- **Document Viewer Comparison**
  Differences between similar documents can be highlighted directly in the viewer.

Workflows

Simply activate the new Filter using the button and press “Apply”



Description		Format	Size	Languages
1.5 Bauordnungsrecht				
1.6 Gebäude (Bewirtschaftung technisch)				
1.7 Mieter				
1.8 Bewirtschaftung (kaufmännisch)				
1.8.1 Bewirtschaftungsverträge				
1.8.2 Versicherungsnachweise				
1.8.2.1 Premium_Invoice_-_Schloßvippach_-_Erfurter_Straße_-_2013-1		.pdf	861.41 KB	
1.8.2.3 Premium_Invoice_-_Schloßvippach_-_Erfurter_Straße_-_2013-1		.pdf	861.41 KB	DEU
1.8.2.4 Insurance_Policy_-_2011-01-18	to confirm	.pdf	375.82 KB	DEU
1.8.2.5 2013-11-25_-_Prämienrechnung_-_Schloßvippach_-_Erfurter_S		.pdf	861.41 KB	DEU
1.8.2.6 2011-01-18_-_Versicherungspolice		.pdf	375.82 KB	DEU
1.8.2.7 Premium_Invoice_-_Schloßvippach_-_Erfurter_Str	to confirm	.pdf	861.41 KB	DEU
1.8.2.8 asdasdasddsa	✕ ✓	.pdf	375.82 KB	DEU
1.8.3 Betriebskostenaufstellung				
1.9 Steuern				
1.10 Rechtsstreitigkeiten				

1.8.2.8 asdasdasddsa

Filter

Group bookmarks

Language
□ FRA □ SPA
□ DEU □ ENG

My findings

Group findings

To approve
Select group

Redaction
□ Original □ Redacted

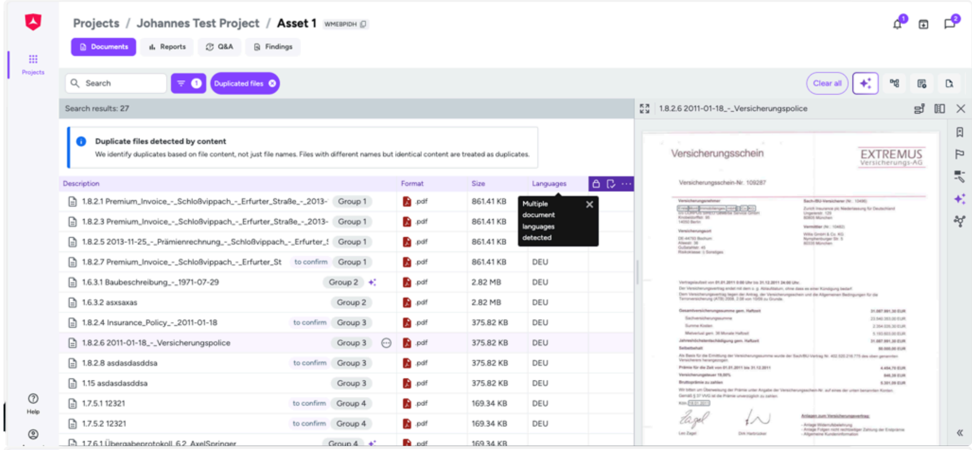
Duplicated files New ●

Uploaded by
Select users

Pages
Select operator | Select number

Reset | Apply

All documents with duplicates inside the dataroom (excluding items in the trashbin) are highlighted. Groups of duplicates are separated by the “Group “ button in the index. Items are identified based on the hashcode despite different document names. Activating properties like “size” indicates duplicates as well.

- 

Projects / Johannes Test Project / **Asset 1** WMEBPIDH

Documents | Reports | Q&A | Findings

Search | 1 | Duplicated files | Clear all

Search results: 27

Duplicate files detected by content
We identify duplicates based on file content, not just file names. Files with different names but identical content are treated as duplicates.

Description			Format	Size	Languages
1.8.2.1 Premium_Invoice_-_Schloßvippach_-_Erfurter_Straße_-_2013-		Group 1	.pdf	861.41 KB	
1.8.2.3 Premium_Invoice_-_Schloßvippach_-_Erfurter_Straße_-_2013-		Group 1	.pdf	861.41 KB	
1.8.2.5 2013-11-25_-_Prämienrechnung_-_Schloßvippach_-_Erfurter_		Group 1	.pdf	861.41 KB	
1.8.2.7 Premium_Invoice_-_Schloßvippach_-_Erfurter_St	to confirm	Group 1	.pdf	861.41 KB	DEU
1.6.3.1 Baubeschreibung_-_1971-07-29		Group 2	.pdf	2.82 MB	DEU
1.6.3.2 asxsaxas		Group 2	.pdf	2.82 MB	DEU
1.8.2.4 Insurance_Policy_-_2011-01-18	to confirm	Group 3	.pdf	375.82 KB	DEU
1.8.2.6 2011-01-18_-_Versicherungspolice		Group 3	.pdf	375.82 KB	DEU
1.8.2.8 asdasdasddsa	to confirm	Group 3	.pdf	375.82 KB	DEU
1.15 asdasdasddsa		Group 3	.pdf	375.82 KB	DEU
1.7.5.1 12321	to confirm	Group 4	.pdf	169.34 KB	DEU
1.7.5.2 12321	to confirm	Group 4	.pdf	169.34 KB	DEU
1.7.6.1 Übergabeprotokoll_6.2_AxelSpringer		Group 4	.pdf	169.34 KB	

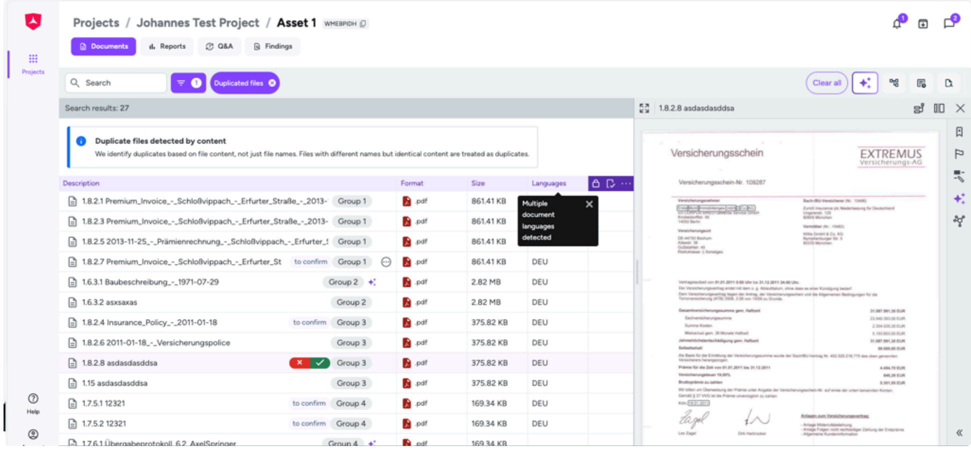
Multiple document languages detected

1.8.2.6 2011-01-18_-_Versicherungspolice

Versicherungsschein

EXTREMUS Versicherungs-AG

Versicherungsschein-Nr: 109287



Projects / Johannes Test Project / **Asset 1** WMEBPIDH

Documents | Reports | Q&A | Findings

Search | 1 | Duplicated files | Clear all

Search results: 27

Duplicate files detected by content
We identify duplicates based on file content, not just file names. Files with different names but identical content are treated as duplicates.

Description			Format	Size	Languages
1.8.2.1 Premium_Invoice_-_Schloßvippach_-_Erfurter_Straße_-_2013-		Group 1	.pdf	861.41 KB	
1.8.2.3 Premium_Invoice_-_Schloßvippach_-_Erfurter_Straße_-_2013-		Group 1	.pdf	861.41 KB	
1.8.2.5 2013-11-25_-_Prämienrechnung_-_Schloßvippach_-_Erfurter_		Group 1	.pdf	861.41 KB	
1.8.2.7 Premium_Invoice_-_Schloßvippach_-_Erfurter_St	to confirm	Group 1	.pdf	861.41 KB	DEU
1.6.3.1 Baubeschreibung_-_1971-07-29		Group 2	.pdf	2.82 MB	DEU
1.6.3.2 asxsaxas		Group 2	.pdf	2.82 MB	DEU
1.8.2.4 Insurance_Policy_-_2011-01-18	to confirm	Group 3	.pdf	375.82 KB	DEU
1.8.2.6 2011-01-18_-_Versicherungspolice		Group 3	.pdf	375.82 KB	DEU
1.8.2.8 asdasdasddsa	✕ ✓	Group 3	.pdf	375.82 KB	DEU
1.15 asdasdasddsa		Group 3	.pdf	375.82 KB	DEU
1.7.5.1 12321	to confirm	Group 4	.pdf	169.34 KB	DEU
1.7.5.2 12321	to confirm	Group 4	.pdf	169.34 KB	DEU
1.7.6.1 Übergabeprotokoll_6.2_AxelSpringer		Group 4	.pdf	169.34 KB	

Multiple document languages detected

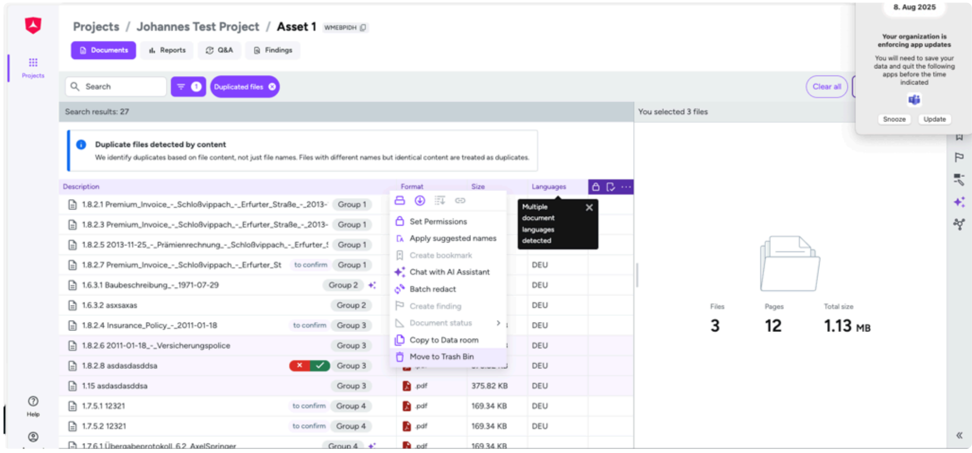
1.8.2.8 asdasdasddsa

Versicherungsschein

EXTREMUS Versicherungs-AG

Versicherungsschein-Nr: 109287

Delete duplicates via multi-selection, right click and “move to trashbin”. Items need to be deleted from the trashbin as well.

- 

Projects / Johannes Test Project / **Asset 1** WMEBPIDH

Documents | Reports | Q&A | Findings

Search | 1 | Duplicated files | Clear all

Search results: 27

Duplicate files detected by content
We identify duplicates based on file content, not just file names. Files with different names but identical content are treated as duplicates.

Description			Format	Size	Languages
1.8.2.1 Premium_Invoice_-_Schloßvippach_-_Erfurter_Straße_-_2013-		Group 1			
1.8.2.3 Premium_Invoice_-_Schloßvippach_-_Erfurter_Straße_-_2013-		Group 1			
1.8.2.5 2013-11-25_-_Prämienrechnung_-_Schloßvippach_-_Erfurter_		Group 1			
1.8.2.7 Premium_Invoice_-_Schloßvippach_-_Erfurter_St	to confirm	Group 1			DEU
1.6.3.1 Baubeschreibung_-_1971-07-29		Group 2			DEU
1.6.3.2 asxsaxas		Group 2			DEU
1.8.2.4 Insurance_Policy_-_2011-01-18	to confirm	Group 3			DEU
1.8.2.6 2011-01-18_-_Versicherungspolice		Group 3			DEU
1.8.2.8 asdasdasddsa	✕ ✓	Group 3			DEU
1.15 asdasdasddsa		Group 3	.pdf	375.82 KB	DEU
1.7.5.1 12321	to confirm	Group 4	.pdf	169.34 KB	DEU
1.7.5.2 12321	to confirm	Group 4	.pdf	169.34 KB	DEU
1.7.6.1 Übergabeprotokoll_6.2_AxelSpringer		Group 4	.pdf	169.34 KB	

Multiple document languages detected

You selected 3 files

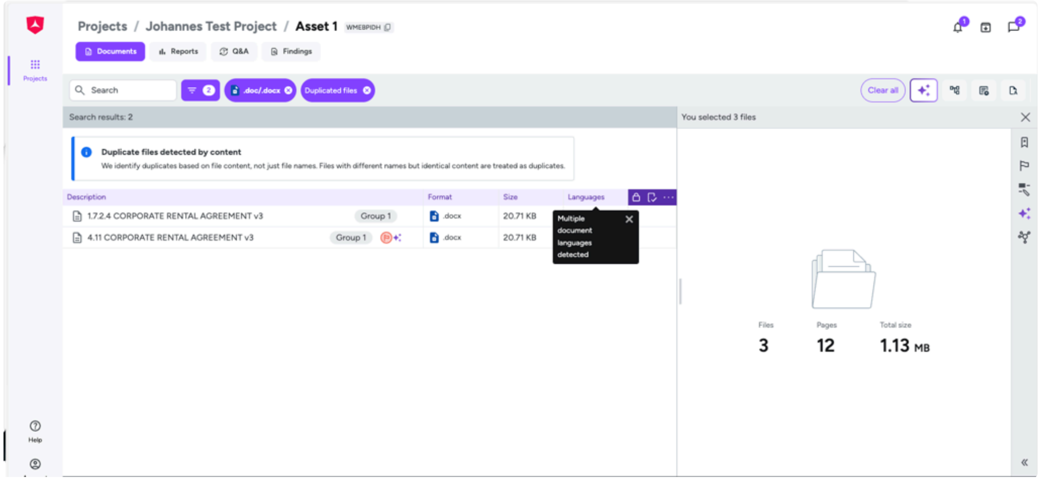
Files	Pages	Total size
3	12	1.13 MB

8. Aug 2025

Your organization is enforcing app updates

You will need to save your data and quit the following apps before the time indicated

The duplicate filter may be easily combined with other filters like the document type



Projects / Johannes Test Project / Asset 1 WMEBPIDH

Search results: 2

Duplicate files detected by content
We identify duplicates based on file content, not just file names. Files with different names but identical content are treated as duplicates.

Description		Format	Size	Languages
1.7.2.4 CORPORATE RENTAL AGREEMENT v3	Group 1	.docx	20.71 KB	
4.11 CORPORATE RENTAL AGREEMENT v3	Group 1	.docx	20.71 KB	

Multiple document languages detected

You selected 3 files

Files	Pages	Total size
3	12	1.13 MB

Technical Implementation

Duplicate detection is based on two complementary approaches:

1. Hashcode Matching

- Each document is assigned a SHA-256 hash at upload.
- If two documents share the same hash, they are exact duplicates.

What is a hashcode?

A hashcode is a unique digital fingerprint generated from a document's content. If two documents have the same hashcode, it means their contents are identical (even if document name/index description or metadata differ).

How it works:

- When a document is uploaded to the dataroom, its **hashcode is calculated via SHA-256 hashing**
    - If another document with the **same hashcode** exists (and is not in the trashbin), it is flagged as a duplicate.
    - If the hashes are different, they are definitely **not duplicates**.
- The duplicate filter queries the database for these hashcode matches and groups them together in the index.
- documents in the **trashbin are excluded** from the duplicate grouping logic.
- Access rights are enforced at the query level, ensuring that users **only see duplicates they are authorized to view**.
- The filter is integrated into the existing filtering framework, allowing multiple filters to be applied concurrently.

2. Content-Based Similarity

- Documents are compared on content level (text and structure).
- Near-duplicates (e.g., two scans of the same contract or signed/unsigned versions) are identified.
- Outputs are ranked by similarity score.

All duplicates are grouped in the index, excluding items in the trashbin.
Access rights are enforced at the query level.

What is content-based similarity?

Content-based similarity measures how closely two documents resemble each other, even if they are **not identical at the binary level**. Unlike hashcodes, which only detect exact duplicates, similarity analysis can recognize near-duplicates such as:

- Scans of the same document with slight differences (e.g., resolution, compression).
- Versions with or without a signature.
- Documents with minor edits or formatting changes.

How it works:

1. **Content extraction**
   - When a document is uploaded, its textual and structural features are extracted (e.g., words, layout, formatting).
2. **Similarity scoring**
   - The system compares these features against existing documents in the dataroom.
   - Each pair of documents is assigned a similarity score (e.g., 98% match).
3. **Duplicate grouping**
   - Documents that exceed a similarity threshold are flagged as near-duplicates.
   - They are clustered and displayed together in the index.
   - Within each cluster, documents are ranked by similarity so that the most relevant matches appear first.
4. **Comparison in viewer**
   - Users can open two documents side by side and highlight differences directly in the document viewer.
5. **Trashbin exclusion**
   - As with hashcode-based detection, documents located in the trashbin are excluded from similarity grouping.
6. **Access control**

    - Access rights are enforced at the query level, so users only see near-duplicates they are authorized to view.

7. **Filter integration**
    - The similarity filter integrates seamlessly with the existing framework and can be combined with other filters (e.g., document type, date, redaction status).

FAQs

FAQs – Duplicate Detection

Q1: How are duplicates identified?
Duplicates are identified either by exact match (hashcode) or by content similarity. Near-duplicates such as different scans of the same contract can also be detected.

Q2: Do renamed documents count as duplicates?
Yes. Renamed documents with identical or highly similar content will be flagged as duplicates.

Q3: Are documents in the trashbin considered duplicates?
No. Trashbin documents are excluded from detection.

Q4: Can I combine the duplicate filter with other filters?
Yes. The duplicate filter works with all other filters (e.g., document type, upload date, redaction status).

Q5: What happens if I delete a duplicate?
When a duplicate is moved to the trashbin, it no longer appears in the duplicate grouping. To permanently remove it, it must also be deleted from the trashbin.

Q6: Will I see duplicates I don’t have access to?
No. Duplicate detection fully respects access rights.

Q7: How are near-duplicates handled?
Near-duplicates are detected using similarity analysis and are displayed ranked by similarity.

Q8: Can I compare duplicates directly?
Yes. Differences can be highlighted directly in the document viewer.

Q9: How fast is duplicate detection?
Hashcodes are pre-calculated at upload. Similarity checks are optimized to work efficiently, even in large datarooms.

Q10: Is this feature available for all datarooms?
Yes. Duplicate detection is integrated into the filtering framework and available across all enabled datarooms..

Tasks

Project Completion Checklist

✅ Before starting each project, make sure you have checked the below common tasks. Kick-Off Call preparation NOT RELEVANT DONE IN PROGRESS	✅ Research and Preparation Phase – Check all points below for completion NOT RELEVANT DONE IN PROGRESS	✅ Ready for development Phase – Check all points below for completion (phase can be overlapped with previous one to speed up and be agile) NOT RELEVANT DONE IN PROGRESS	✅ In Work – Check all points below for completion, this includes Rollout preparation by PMs NOT RELEVANT DONE IN PROGRESS
☑ Create an Epic in Jira, link it to this project space	☑ Extend competitor research and complete Competitor, Market, Trend research Template	☑ QA ticket estimates	☑ All BE tickets completed
☑ Set estimated Start and End dates of Epic for Research Preparation status	☑ Define clear project requirements for design and developers make use of Jobs to be done or simple user story framework	☑ QA testing ticket creation	☑ All FE Tickets completed
☑ Create Project Documentation page in Confluence	☑ Create User flows and link Miro board	☑ FE and BE Technical task creation was completed	☑ QA testing completed
☑ Add general background information	☑ Gather relevant data on time, define metrics in exchange with Developers and stakeholders	☑ FE estimated all User stories	☑ Marketing and Design briefing prepared (can be done earlier as well)
☑ Add all links, including feature requests	☑ User Journey and final flow (Lo-fi/Wireframes) finalized	☑ initial Release date estimate available	□ Alignment and final plan for marketing activities and campaigns
☑ Complete the Product Brief Template	☑ Technical Research by developers completed	☑ Inform Help Center team enough time in advance, in case visual changes will happen that would affect their content.	☑ Feature announcement in Product announcement prepared and developed

☑ Plan a kick-off call including all features leads, CS Leads, and other relevant stakeholders (Marketing, Finance, Sales) and bring project docs and product brief as the foundation	☑ Hi-fi/final designs created	☑ have a regular status check and update exchange with stakeholders	☑ Effect on guided tours, updated or new one created, tested, code added from devs and released
☑ Plan recurring meetings with the core team and extended stakeholder team in the kickoff call	☑ Milestone meeting with all project involved people and relevant stakeholders and team leads		☑ Training material created and published in Confluence
☑ Get rough estimate by Design, BE, FE regarding project effort, feasibility and resource outlook and add it to corresponding Discovery page.	☑ All Assets created by Design		☑ Training session planned and completed with Sales and CS
☑ Create a project chat involving only key stakeholders (PM, BE, FE, QA, DES, CSM)	☑ Write all user stories		☐ Release notes prepared
	☑ Add final links to the Epic and confluence project page (miro, figma...)		☐ Help Center article created or updated with CS
	☑ QA review of the tickets		☐ Marketing and relevant Design Material for Release announcements ready
	☑ Initiated pricing strategy with finance (when applicable)		☑ Pricing strategy is ready for release, finance articles created, contract updated
	☑ Initiated legal evaluation, impact and task (when applicable)		☑ Legal adjustment are ready for release like T&C adjustments
	☑ Confident with Epic scope: Review epic for first release and evaluate value compared to effort, is MVP fine or do we need more value or iterations		☑ Website adjustment are ready for release
			☑ Included in release candidate: add <release candidate number> here

			☑ Finalize documentation and close relevant tickets (DBT, RICE)
			☑ Follow next iterations based on your priorities