MIGRATIE VAN GROTE DATASETS

Research Drive training
21 april 2021

Training van vandaag

- Inhoud:
  - 10:00 – 10:30: inleiding en presentatie
  - 10:30 – 11:00: demo gebruik rclone
  - 11:00 – 11:30: vragen en discussie

- Presentatie wordt opgenomen
- Verwijzing naar Research Drive wiki:

https://wiki.surfnet.nl/display/RDRIVE/SURF+Research+Drive+wiki

Vandaag *niet* behandeld

- Casussen:
  - Kleine hoeveelheden data op eigen laptop of elders
  - Migratie via federatieve shares
  - Transfers via de browser

  https://wiki.surfnet.nl/display/RDRIVE/How+to+upload+or+download+your+files
- Tools en applicaties:
  - Tranferapplicaties zoals FileZilla, WinSCP
  - cURL
  - Integratie in eigen applicaties
  - Batch / scripting

Voor we beginnen

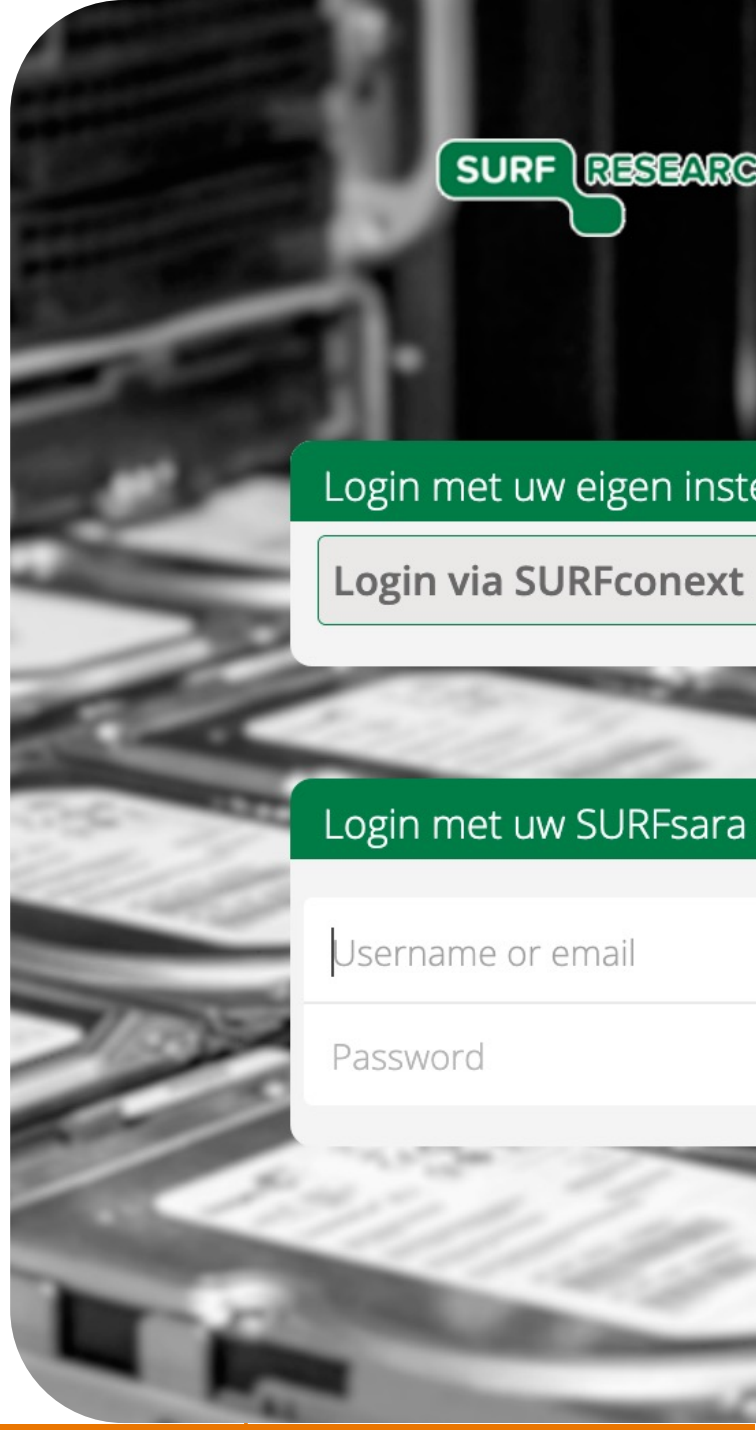
- Datatransfers van en naar Research Drive

- Benodigdheden:
  - Internetverbinding (beetje snel graag)
  - Toegang tot Research Drive
  - WebDAV credentials
    https://wiki.surfnet.nl/display/RDRIVE/How+to+upload+or+download+your+files#Howtouploadordownloadyourfiles-GettingyourWebDAVcredentials
  - Toegang tot te migreren data
- Verder:
  - Inzicht in huidige en benodigde structuur

Probleem

- Data opgeslagen in verschillende locaties buiten Research Drive
- Onoverzichtelijk en geen inzicht op inhoud
- Data niet toegankelijk of offline opgeslagen
- Toegang verschaffen (delen, beheer) is niet mogelijk of lastig
- Versiebeheer van bestanden lastig te regelen
- Datavolume en aantal bestanden

Hoe krijg ik de vrij toegankelijke data van mijn instituut in Research Drive?



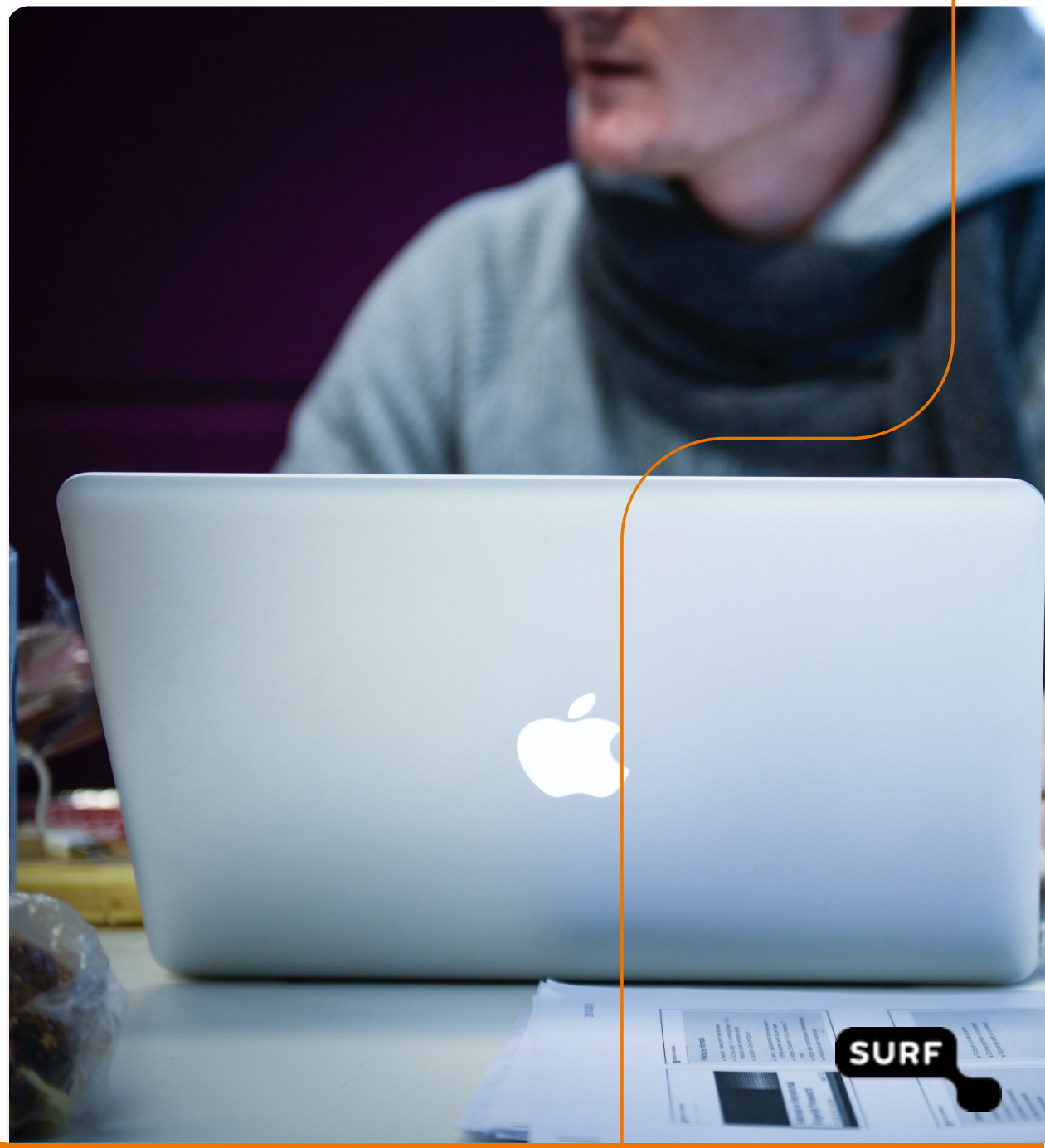
Intakeformulier

- Waar staan de bronbestanden voor migratie naar RD?
- Globale karakteristiek van de te migreren bestanden
- Is opschoning van de te migreren bestanden gewenst?
- Blijft de mappenstructuur van de te migreren bestanden behouden?
- Blijft de rechtenstructuur van de te migreren bestanden behouden?
- Worden de bestanden gemigreerd naar een actieve projectruimte?
- En:
  - Totaal datavolume
  - Aantal bestanden
  - Frequentie

Jurgen Mollema (HU)

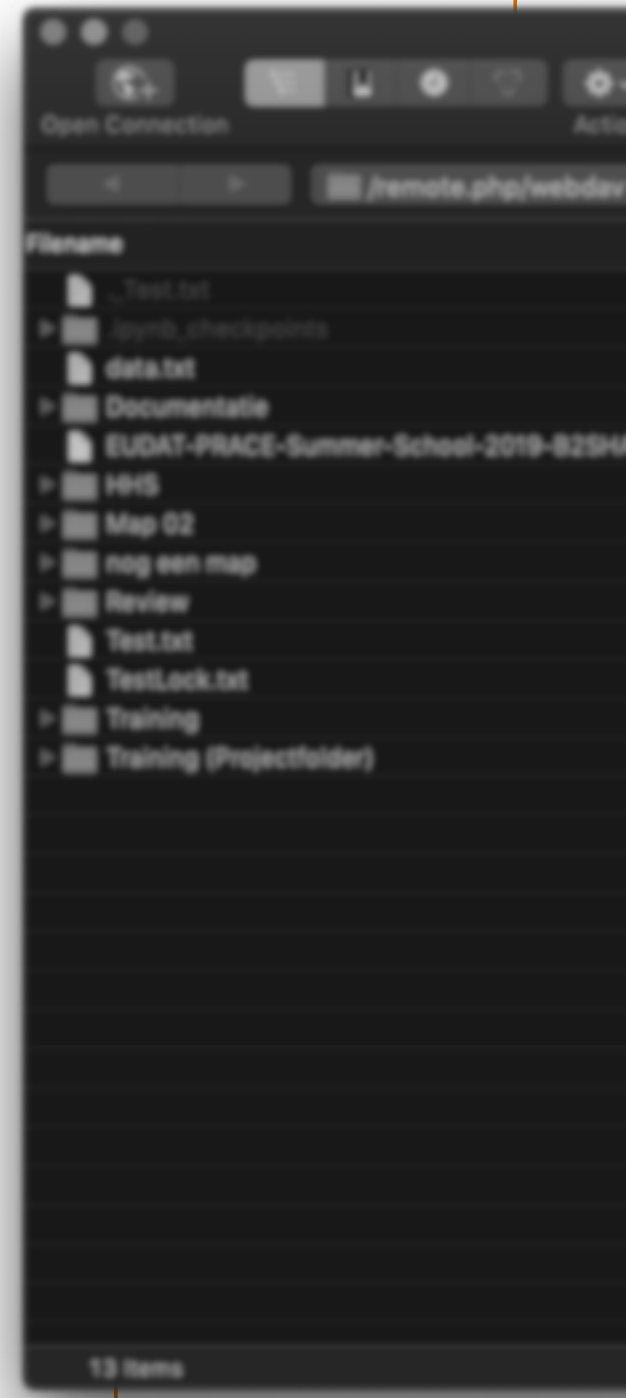
Uitdaging

- Hoe krijg ik al mijn data in Research Drive:
  - Zo snel en efficient mogelijk?
  - Zonder overschrijving van bestaande data?
  - Met minimale inspanning?
  - Vanaf verschillende locaties?
- En:
  - Behoud van structuur?
  - Met behoud van eigenaarschap en/of rechten?
  - Herhaling?

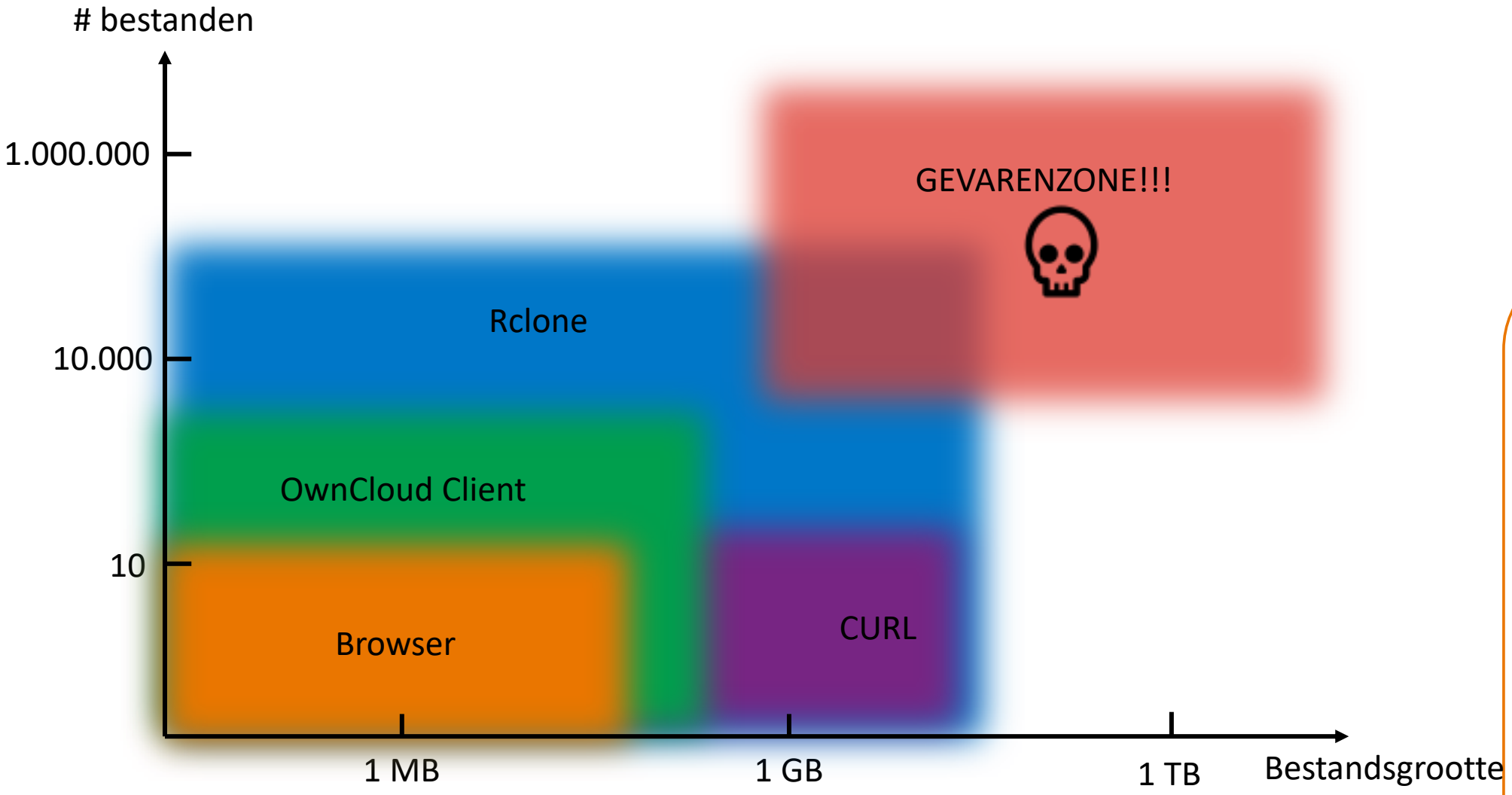


Transfer/synchronisatie tools

- Applicaties (WebDAV ondersteuning):
  - Owncloud client
  - WinSCP
  - Cyberduck
  - Veel andere
- Voordelen:
  - Makkelijk(er) in gebruik
  - Overzichtelijk
  - Drag ‘n’ drop
- Tools:
  - rclone
  - cURL
  - Rsync
- Voordelen:
  - Batch / scripting / automation
  - Flexibel(er)
  - Testen
  - Logging



Wanneer gebruik ik welke tool? https://wiki.surfnet.nl/display/RDRIVE/How+to+upload+or+download+your+files



Rclone

- Tool voor beheer van bestanden lokaal en in de cloud

- Alleen bruikbaar via command-line*!

- Open source: broncode volledig inzichtelijk: https://github.com/rclone/rclone
- Parallelle transfers: meerdere bestanden tegelijkertijd versturen
- Synchronisatie: twee locaties up-to-date houden met elkaar
- Mounting: lokaal bestanden toegankelijk maken die remote zijn opgeslagen

- Vele back-ends ondersteund, waaronder OwnCloud (Research Drive, SURFdrive), SFTP (Data Archive, Cartesius, Lisa)

Rclone: belangrijk

- Alleen eenrichtingsverkeer! (A => B)
- Werken met lokaal/remote, remote/remote of remote/lokaal
- Incomplete bestanden moeten volledig opnieuw worden geupload
- Gebruik specifieke opties om bestandselecties te maken of vergelijkingen in te stellen bij synchronisatie

https://rclone.org/faq

Installatie

- Rclone hoeft niet geïnstalleerd te worden, kan wel!
- Gebruik je package manager
- Of download rclone: https://rclone.org/downloads/
  - Selecteer de juiste versie voor je OS!

- Of gebruik de command line (MacOS/Linux):

  curl https://rclone.org/install.sh| sudobash

https://rclone.org/downloads

Commando’s uitvoeren

- Gebruik:
  - Windows: command prompt
  - MacOS: terminal
  - Linux: terminal

- Uitvoeren:
  - rclone <opties> <taak> <bron> <doel>
  - Bijv: rclone –vP copy bestand.txt RD:bestand.txt

Unieke naam remote: ‘RD’

https://rclone.org/docs

Wat is een remote?

- Locatie van opgeslagen data (lokaal of cloud)
- Heeft (indien in de cloud):
  - Unieke naam
  - Opslagtype
  - Een of meer opties afhankelijk van opslagtype
  - Vaak authenticatie

- Toevoegen aan configuratie voor gebruik!

Configuratie

- Commando: rclone config
- Toevoegen nieuwe ‘remotes’
- Benodigd voor SURFdrive/Research Drive:
  - Unieke naam
  - Storage type (WebDAV)
  - URL (e.g. https://researchdrive.surfsara.nl)
  - Service type (OwnCloud)
  - WebDAV gebruikersnaam
  - WebDAV wachtwoord

  https://wiki.surfnet.nl/display/RDRIVE/How+to+upload+or+download+your+files#Howtouploadordownloadyourfiles-GettingyourWebDAVcredentials
- Geavanceerde configuratie overslaan!

Interessante opties tijdens uitvoering

Optie	Beschrijving
-v	Extra informatie weergeven tijdens transfers
-vv	Nog meer informatie weergeven
-vvv	Nóg meer informatie weergeven
-P	Voortgang per transfer weergeven
-n	“Dry-run”: test het commando zonder het daadwerkelijk uit te voeren
-q	Zo min mogelijk weergeven

- Combineer opties: -vPnq
- Alle opties weergeven: rclone help flags
  - Dit zijn er meer dan 500!

Interessante opties voor synchronisatie

Optie	Beschrijving
-u --update	Alleen niet-bestaande of oudere bestanden vervangen
--ignore-existing	Alleen niet-bestaande bestanden toevoegen
-f --filter	Filter voor bestandsnamen
-I --ignore-times	Alles (opnieuw) synchroniseren
--ignore-size	Bestandsgrootte negeren
-c --checksum	Bestandsdatum negeren
--ignore-checksum	Checksum niet vergelijken

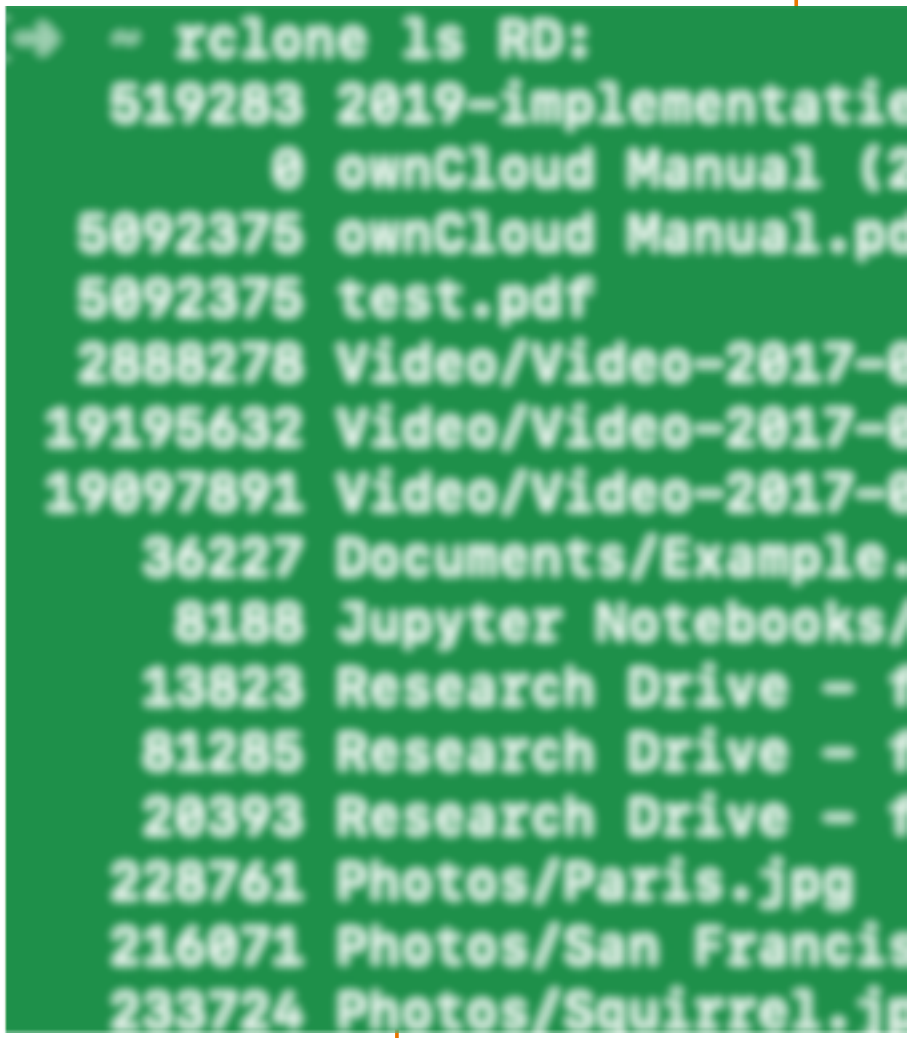
- Combineer opties: -ucf <filter>
- Alle opties weergeven: rclone help flags
  - Veel specifiek voor type storage van remote

Remote testen / weergeven

- Commando: rclone ls <remote>:
- Laat achter elkaar alle bestanden van remote zien

- Opties:
  - Alleen hoofdfolder: gebruik --max-depth 1

- Alternatieven:
  - Alleen directories: rclone lsd <remote>:
  - Met datum: rclone lsl <remote>:

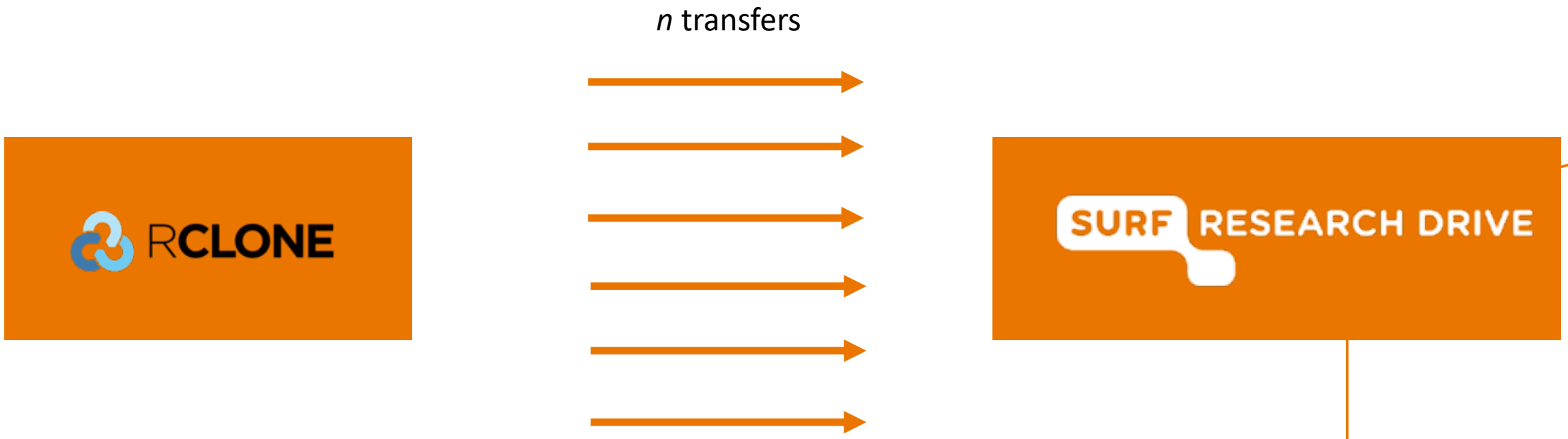


Belangrijkste commando's

- Help weergeven:
  rclone help
- Configuratie:
  rclone config
- Enkel bestand kopiëren:
  rclone copyto <bestand> <remote:bestandsnaam>
- Kopiëren (lokaal naar remote):
  rclone copy <bestanden/folder> <remote:folder>
- Kopiëren (remote1 naar remote2):
  rclone copy <remote1:bestanden/folder> <remote2:folder>
- Synchroniseren:
  rclone sync <remote1:bestanden/folder> <remote2:folder>

https://rclone.org/docs

Schema (lokaal naar RD)



- Commando: rclone copy bestanden* RD:gebruiker1/

https://wiki.surfnet.nl/display/RDRIVE/Client-side+copy

Belangrijke opties

- Verwerking kan lang duren: gebruik --timeout <tijd>
- Stel het aantal parallelle transfers in: optie --transfers <n>
- Versnel authenticatie: gebruik --use-cookies

- Vuistregels:
  - Bestanden < 5 GB: 24 transfers
  - Bestanden > 5 GB: 100 / (grootste bestanden GB) = *n* transfers (max 24)
  - Time-out instellen: grootste bestand in GB x 10 minuten: --timeout 100m
  - Bij veel kleine bestanden in dezelfde directory: altijd --use-cookies gebruiken

https://wiki.surfnet.nl/display/rdrive/rclone+performance+considerations

Performance vs. bestandsgrootte en parallellisatie

- De invloed van structuur van bestanden en parallellisatie op de duur van de migratie

Totaal volume	Bestands-grootte	Aantal bestanden	Tijd (s, 1 tr.)	Tijd (s, 24 tr.)
50MB	10kB	5000	1300	350
50MB	100kB	500	140	45
50MB	500kB	100	29	25
50MB	1MB	50	17	15
50MB	2MB	25	11	10
50MB	5MB	10	10	10
50MB	10MB	5	10	10
50MB	25MB	2	15	15

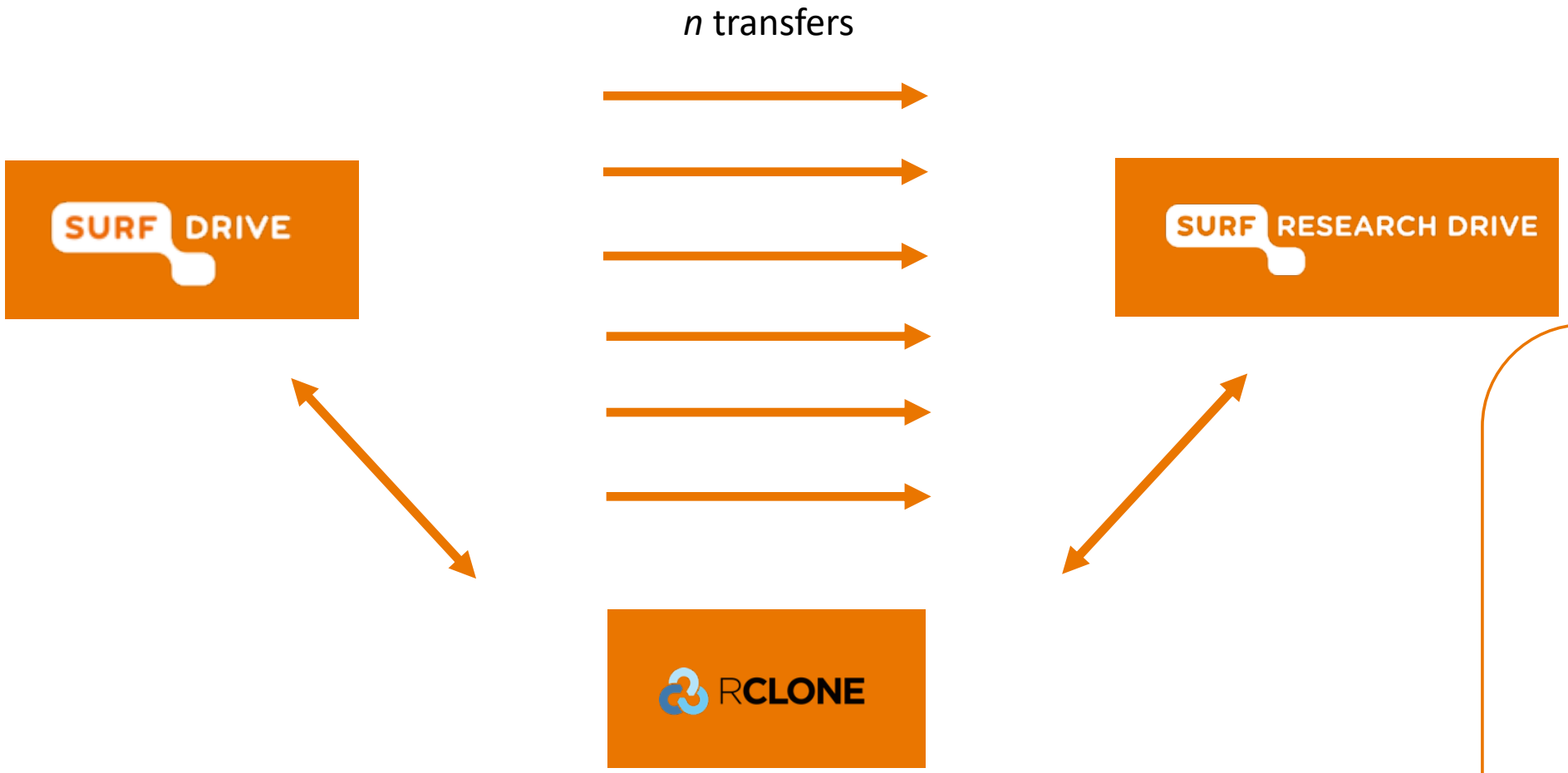
https://wiki.surfnet.nl/display/rdrive/rclone+performance+considerations

Schema (remote naar RD)

- Commando: rclone copy SD:folder1/ RD:folder1/

https://wiki.surfnet.nl/display/RDRIVE/Client-side+copy

Schema (‘server-side copy’, remote naar RD gefedereerd)



- Commando: rclone copy RD:folder1/ RD:folder2/

  - https://wiki.surfnet.nl/display/RDRIVE/Share+with+an+another+Research+Drive+instance
  - https://wiki.surfnet.nl/display/RDRIVE/Server-side+copy

Keuzematrix

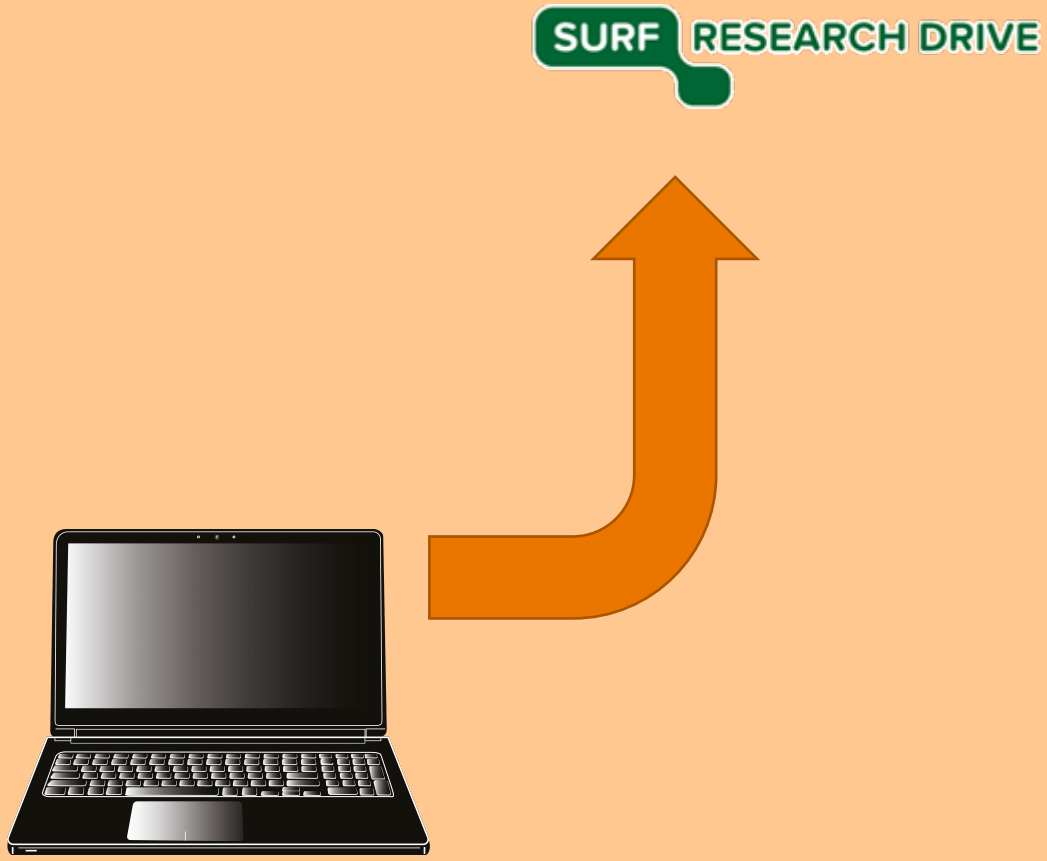
Bron	Datavolume	Aantal bestanden	Aanbevolen tool voor migratie
SURFdrive / Research Drive	**< 1 GB**	**< 500 files**	OwnCloud desktop client
	< 1 GB	**> 500 files**	OwnCloud desktop client Rclone (client-side copy)
	> 1GB	**< 500 files**	Rclone (server-side copy)
	> 1GB	**> 500 files**	Rclone (server-side copy)
Lokaal (laptop/network drive)	**< 1 GB**	**< 500 files**	OwnCloud desktop client
	< 1 GB	**> 500 files**	OwnCloud desktop client
	> 1GB	**< 500 files**	Rclone
	> 1GB	**> 500 files**	Rclone
Remote server (Onedrive, Dropbox, …)	**< 1 GB**	**< 500 files**	Rclone
	< 1 GB	**> 500 files**	Rclone
	> 1GB	**< 500 files**	Rclone
	> 1GB	**> 500 files**	Rclone

https://wiki.surfnet.nl/display/RDRIVE (TBD)

Demo Datamigratie

1. **Configuratie**
   - Instellen rclone
2. **Verkenning**
   - Verbinden met de remote
3. **Performance**
   - Parallelisatie
   - Impact bestandsgrootte
4. **Client-side copy**
   - Migratie van remote naar remote, via lokale machine
5. **Server-side copy**
   - Migratie van remote naar remote, direct
   - Toepassing, performance, nadelen

Lokale bestanden overzetten



Client-side migration

SURF DRIVE

SURF RESEARCH DRIVE

Server-side migration

SURF DRIVE → SURF RESEARCH DRIVE

HAPPY MIGRATING!

SURF Research Services

servicedesk@surfsara.nl

https://www.surf.nl

@SURF_NL