Nyttan finns i användandet!

Projektrapport SwePub: Datakvalitet ur ett lärosätesperspektiv

Kristoffer Antonsson, Kristin Olofsson, Anders Söderbäck, Daniel Wadskog

Sammanfattning

Projektgruppen har, utifrån dialog med lärosäten, VR och KB, sett ett stort intresse och en stor potential i ett utvecklat SwePub. Samtidigt finns brister i trovärdigheten och en rädsla för att en investering i SwePub kommer visa sig meningslös. Vi ser ett problem i att diskussionen om SwePub handlar om brister i datakvalitet och den potentiella kostnaden i att åtgärda dessa. Istället måste vi börja **använda** SwePub, och prata om den **nytta** som uppstår i en sådan användning.

För att åstadkomma detta rekommenderas en basnivå för leverans till SwePub som bör vara genomförbar för lärosäten utan större investering samtidigt som den gör det möjligt för VR att börja använda SwePub. Vidare rekommenderas att SwePub lyfts av både KB, VR och SUHF som en långsiktigt prioriterad tjänst gentemot högre utbildning och forskning.

Om detta inte kan realiseras, och SwePub inte börjar användas, rekommenderas istället att SwePub så snart som möjligt avvecklas.

I projektuppdraget har vi tagit fasta på

Identifiera och föreslå åtgärder, både i SwePub och på lärosätena, för att höja kvaliteten på innehållet i SwePub. Föreslagna åtgärder ska vara **resursuppskattade** och **prioriterade**.

Åtgärderna bör utgå från ett **helhetsperspektiv** kring hanteringen av publikationsdata, men avgränsas till att gälla kvalitetshöjning av publikationsdata.

Arbetet ska inte utgå från befintliga system, utan **i första hand titta på behov, nytta och processer** vid lärosätena.

Nyttan med SwePub behöver beskrivas på ett sätt som **gör lärosätena delaktiga i arbetet**.

Svårigheter med resursuppskattning

Feltyper, system och arbetsprocesser skiljer sig åt mellan olika lärosäten och därför har en enhetlig resursuppskattning varit svår att genomföra.

Istället för en detaljerad resursuppskattning har vi valt att lägga vår rekommendation på en nivå vi bedömer är fullt möjlig för samtliga lärosäten att uppnå inom nuvarande organisation, utan större insatser eller att extra medel tillförs.

Mellan januari och mars 2018 har genomförts

- Avstämningsmöten med KB (Tuija Drake, Martin Malmsten och Bodil Wennerlund) och Vetenskapsrådet (Henrik Aldberg och Elisabeth Sjöstedt)
- Två workshops med 50 deltagare från 30 leverantörsorganisationer
- Fyra fysiska projektmöten samt ett flertal virtuella möten
- Fördjupad analys av datakvalitet och arbetsprocesser samt av VR:s beställarunderlag

Resultat från workshops med lärosätena*

- Tydligt önskemål om stöd för datakvalitet (riktlinjer, validering, support från KB)
- Stort intresse för att själva kunna återanvända data från SwePub
- Intresse för hur Swepub-datat kommer att användas
- Viss tvekan ”Blir det något den här gången?”

* Fotnot: Termen lärosäten inbegriper alla levererande organisationer. En mer utförlig redovisning av workshoparna: https://docs.google.com/document/d/16rZb7nz1tGfgrB86IeBP-OPEwaCo8YtSDz-EeX6dqvk/edit?usp=sharing

Tre framtida scenarier för SwePub

1. De insatser som görs initialt ger tillräcklig kvalitetsökning för VR och är tillräckligt enkla för medlemmarna att bidra till för att kunna realiseras snabbt
    a. Lärosätena prioriterar kvalitetsarbete utifrån en nedbantad version av VRs beställarunderlag
    b. KB avlastar lärosäten och andra anslutna organisationer med automatisering
2. Stor satsning på att få fullständig kontroll på publikationsdata inom SwePub, SwePub blir en “megadatacirkulator”
    a. Alla poster är deduplicerade och unikt representerade i SwePub
    b. SwePub fungerar väl som en nod i ekosystemet för forskningsinformation och möjliggör återanvändning av publikationsdata i många system
    c. Stora möjligheter att återanvända SwePub-data maskinellt
3. SwePub läggs på is
    a. SwePub klarar inte av att få tillräcklig kvalitet i data för analys eller att visa på nyttan i tjänsten.
    b. Utvecklingen stagnerar och anslutna organisationer jobbar inte aktivt med datakvalitet

Scenario 1: SwePub börjar göra nytta

I detta scenario börjar SwePub göra faktisk nytta genom att användas, åtminstone av Vetenskapsrådet. Ett sådant nyttiggörande av SwePub kommer också att göra lärosätenas kvalitetsarbete meningsfullt. För att uppnå detta behöver följande åtgärder genomföras:

- Implementera en basnivå för dataleveranser till SwePub (sida 9-12)
- Vetenskapsrådet börjar använda datat i SwePub för analys (sida 13)
- Lyft, tydliggör och förankra en långsiktig vision för SwePub (sida 15)

En beskrivning av nyttan med Scenario 1 återfinns på sida 14.

Scenario 1: Basnivå för leverans till SwePub

Denna basnivå är en kompromiss mellan VRs krav på kvalitetssäkrad data för analys, KBs möjligheter att stödja medlemmar med automatisering och lärosätenas nuvarande kapacitet för kvalitetsarbete. Det vill säga: *Rimliga analyser till rimlig insats.*

	Basnivå	**Motivering**
Ämnesklassning	KB ansvarar*	KB har redan flaggat att de vill försöka automatklassa ämnen i SwePub
ISSN-lista	KB ansvarar	KB bör validera data och ha en del egna kritiska aouktoritetsregister
ISSN, ISBN	Ja	ĸopplar mot DOAJ och nordiska/norska listan
DOI	Ja**	**Enbart för poster som har DOI, det är en viktig identifierare.
Totalt antal upphovspersoner	Ja	Viktigt för fraktionerade analyser, borde inte vara svårt för lärosäten
Egna affilieringar	Ja	Det ger nytta även för det enskilda lärosätet att ha dessa korrekta

Scenario 1: Strykningar från VR:s beställning

Vi menar att dessa önskemål i nuläget kan strykas från VR:s beställarunderlag. Detta eftersom de tillför marginellt värde och/eller tar stora resurser i anspråk på lärosätena.

	Basnivå	**Motivering**
Kön	Nej	Levereras ej pga dataskyddslagstiftning och stora insatser lokalt för att kunna leverera.
Externa affilieringar	Nej	Det blir ett jättejobb att få fram information om allting utöver egna organisationens upphovspersoner från lärosätena. Ska detta göras måste det göras centralt.
Outputklassificering	Nej	Oklart värde, vi väntar tills basnivån är implementerad och ser om detta behövs.
Lokala personID:n	Nej	Inget omedelbart behov för VR:s analyser, behövs framåt, satsa på ORCID.
DOI eller unik SwePub-identifierare på alla publikationer	Nej	Alla poster har inte DOI i original och KB har svårt att tänka sig unika SwePub ID:n inom kort.

Scenario 1: Arbete på lärosäten

- Ett antal **prioriterade** städinsatser lokalt. Dessa bör initialt röra feltyper kopplade till ISSN, ISBN, DOI, antal upphovspersoner.
- Viktigt att påpeka att det handlar om ett relativt litet antal poster (ca 10%*).
- Den initiala städningen löser så klart inte allt, men är det som är prioriterat utifrån VR:s beställarunderlag.
- Lärosätena behöver ta ett tydligt ansvar för att leverera kvalitativ data enligt föreslagen basnivå.
- I övrigt anser vi att rekommendationerna för leverans till SwePub bör fortsätta att gälla.

* Siffran är inte helt lätt att uppskatta, men bygger på att de största feltyperna ignoreras (namnvarianter ca 50%, PersonID saknas ca 10%) eller löses centralt (ämne saknas ca 30%). Procenten anger andel av poster markerade med bristfällig metadata i SwePub.

Scenario 1: Stöd från KB

- Representationen av datakvalitet på http://bibliometri.swepub.kb.se/databearbetning behöver korrigeras. Den redan relativt goda kvaliteten bör visas upp, istället för bristerna i datat.
- Inför stöd för automatisk ämnesklassificering.
- Utveckla en bra valideringsfunktion som direkt visar fel vid leverans.
- Förtydliga nationella riktlinjer; särskilt fokus bör läggas på att lösa frågor som genererar bristande datakvalitet.
- Utarbeta hållbara förhållningssätt till brister som orsakar fel men som inte går att åtgärda med städning av data (t.ex. hur hantera serier utan ISSN?).
- Bättre support till lärosäten vid leverans (inklusive manuell ”support på plats”).
- Vidareutveckla de arbetsflöden som rör ”dubblettkandidater” och ”samarbetspublikationer”.
- Utred och förhandla hur WoS-id:n kan användas för cirkulering inom SwePub

Scenario 1: Insats från VR

- VR behöver börja göra analyser på data från SwePub
- Arbetet kan initialt bedrivas i form av test, men avsikten måste vara att inom en relativt snar framtid börja använda SwePub i skarpt läge
- Resultatet av VR:s analyser behöver göras synligt för lärosätena för att ge faktiskt återkoppling på lärosätenas kvalitetsarbete

Scenario 1. SwePub - basnivån och nyttan

- Mer heltäckande källa för nationella analyser och benchmarking. SwePub täcker upp då publikationer inte finns i Web of Science. Speciellt viktigt för Hum/SAM-området
- Det går att börja använda data från SwePub för analyser
- Benchmarking. Jämförelser med andra lärosäten, publikationer per ämne o.d.
- Central automatisk ämneskategorisering (SCB/UKÄ) sparar resurser lokalt på lärosäten och blir mer homogen.
- Lättanvända gränssnitt/funktioner underlättar det lokala städ- och kvalitetsarbetet
- Om SwePub används för analys kommer täckningen och kvaliteten att öka successivt

Scenario 1: Långsiktig vision för SwePub

Leveranser på basnivå är enligt vår analys antingen ett nödvändigt första steg på en långsiktig satsning, eller en kortsiktig livsuppehållande åtgärd innan ett fortsatt förfall. En långsiktig vision för SwePub behöver lyftas och förankras hos både VR, KB och lärosäten. Denna vision måste motivera samtliga berörda intressenter att göra de investeringar som krävs för att få SwePub att fungera. För att genomföra detta krävs att

- KB tydligt visar att SwePub är en långsiktigt prioriterad verksamhet gentemot högre utbildning och forskning.
- KB, VR och SUHF för en gemensam dialog om den fortsatta utvecklingen av SwePub och den nytta SwePub kan göra för samtliga parter. KB behöver ta ansvar för denna dialog.

Scenario 2: SwePub som “megadatacirkulator”

Genom fortsatt utveckling efter Scenario 1 har SwePub stora möjligheter att göra än mer nytta på såväl lokal lärosätesnivå samt nationell nivå. SwePub fungerar i detta scenario som en nationell nod för att cirkulera data mellan flera system. En sådan utveckling kräver större investeringar, och det måste därför finnas tilltro till långsiktigheten i utvecklingen av SwePub. Detta scenario kräver därför att Scenario 1 först är genomfört.

Nyttan med ett kraftigt vidareutvecklat SwePub beskrivs på sida 17-18.

Scenario 2: SwePub - megadatacirkulator och nyttan

- Central databerikning från WoS, Scopus, PubMed
- SwePub kan berika och kvalitetsförbättra datan maskinellt (auktoritetsregister, homogenisera adresser, id:n etc.)
- Erbjuda bra exporter och API:er så att lärosätena kan återanvända det berikade datat lokalt
- Lärosätena kan få tillgång till en forskares publicering vid andra lärosäten i Sverige
- Nordiska listan integreras med Swepub. Efterlängtat och användbart för lärosätena.
- Ingående poster är deduplicerade vilket ger tillförlitliga analysresultat

Scenario 2: Swepub - megadatacirkulator och nyttan

forts.

- Elegant integrering med Prisma som kan underlätta för forskaren att ange relevanta publikationer vid ansökan. Kräver gediget UX-arbete.
- Flexibel klustring av forskningsoutput som underlag för finansiärer och i strategi- och forskningspolitiskt arbetet
- Aktivt arbeta för att forskarna skaffar ORCID, som en ersättning/komplement till person:id
- Koppling Swepub - SweCRIS. Stöd för VR och finansiärer att följa upp deras OA-policy. Få tillbaka projektdata för de som hanterar projektinformation lokalt.

Scenario 3: SwePub läggs på is

Om SwePub inte klarar av att nå en datakvalitet som är tillräcklig för analys, eller inte klarar av att visa tillräcklig nytta i användningen av det data som finns, kommer utvecklingen att stagnera. Detta kommer i sin tur leda till att anslutna organisationer slutar jobba aktivt med datakvalitet, varpå kvaliteten sjunker ytterligare och nyttan blir än svårare att uppnå. Om SwePub inte kan användas påbörjas istället en långsam död.

Projektgruppen tror det finns stora vinster i att undvika detta och istället satsa på en hållbar utveckling av SwePub enligt Scenario 1 och 2.. Att KB, VR och lärosätena påbörjar ett nytt utvecklingsarbete utan långsiktig trovärdighet bedömer vi dock vara ett enormt slöseri med offentliga medel.

I det fall Scenario 1 inte trovärdigt kan genomföras rekommenderar vi istället ett medvetet fattat beslut att avsluta arbetet med SwePub.

Rekommendation

Projektgruppen rekommenderar att arbete påbörjas med att utveckla SwePub enligt den basnivå för leveranser som beskrivits i Scenario 1.

VR rekommenderas att så snart som möjligt inleda arbete med analyser baserade på SwePub som beskrivits i Scenario 1.

KB rekommenderas att ansvara för att på ett trovärdigt sätt lyfta den långsiktiga visionen för SwePub. Detta arbete bör ske tillsammans med VR och SUHF.

Om ovanstående ej kan genomföras rekommenderas istället KB att ansvara för ett snabbt avvecklande av SwePub.