Alfresco TTL #167 Script

Transform your content into AI-ready data with the Knowledge Enrichment APIs

Speakers:

- Angel Borroy, Developer Evangelist, Hyland
- Nabih Metri, Product Manager, Hyland
- Hein Ragas, Product Manager, Hyland

Angel Borroy: Welcome, to everyone, to this new Alfresco Tech Talk, live. This is the one hundred and sixty, seventh edition. And as usual, I'm going to start, by, reviewing some news.

I will present later our speaker. And the first thing is just to review some news about the community. So, these are, like, projects, from the community and also projects from our laboratories that are relevant in the last month since we, hosted the previous TPL. So the first one, is the spin AI model comparison. So as you know, Docker released on the last month the Docker model runner.

So this Docker model runner is able, to run locally elements. That is more or less the same, feature provided by Ollama in the previous, experiments we did with that. So I was just, creating a project to compare how this was performing and how hard was, to switch from Ollama to Docker model runner. By now, Docker model runner is only available for, Mac with silicon chips, but, in the next month, it will be available also for Windows. So you can take a look, but, the first impression is that switching to Docker model runner should be, like, a natural, movement in the in the next month.

Also, something that is relevant is that this Docker model runner is not running inside the Docker engine. So it's running natively in your computer. So the performance is as high as Ollama. So it's more or less the same thing. So, this is the conclusion of the of the project.

Right? So you're gonna skip the project. In any case, we have also some other projects like this one from Sidon seventy five that is, onetime password, for two factor authentication for Alfresco. You have the the link. There is also an experiment to store the Alfresco embeddings, with, Qdrant.

Remember that you can use elastic search as a vector database or any other, but in this case, this one is providing the storage with quadrant. Also from our colleagues of, at OCD, we have a new, review for the Alfresco audit, sir. So this is, statistics add on for Alfresco, sir, very popular. So there is a new version for this module, and there is also, documentation and notes by our colleagues, Francesco Papini, that was taken while, they are just evolving the, on premise deployment of, of our first going in their, with their customers and in their, projects. That is also really a nice reading.

Also, we had some contributions. So for the Alfresco Docker installer, the,

deployment for releases 7.x was broken, but our colleagues, you know, provided, sorry for for the pronunciation, provided the the fix for that. So for now, that is again working. There was also some, some reports of the office mobile workspace, the later release one eleven. Had an issue.

So it was not able to connect by default. So we are working in that, but there was also people in highlight connect and also, Sergio, from the Discord server. Yes. Reporting, this problem. So we are starting with the with the resolution of that.

There was also a relevant vulnerability. This, CV, two zero two five, two four eight one three. That is a Tomcat vulnerability. There is no impact in our first call products. But, again, thanks for, thanks for reporting this.

We will, we will be providing the official statement, but, you are safe. I mean, there is no no problem with this specific vulnerability. And, finally, two new blog post, related with this local model runner. Again, just trying to to switch from one to, to another model and making some experiments with it. Also, a new announcement.

So finally, it's open. So CommunityLIVE in US, for this year, it's open. So you can register in the conference. It will be, it will be happen in Las Vegas, August twenty five to twenty eight, and there is a new initiative from Hyland that is a developer track. So the the first thing so you can just we will be including more activities and more information with the time.

But, from now, you can see that there will be a hackathon, and there will be on the training specific training for developers. It will be also related with the with the concepts and products that Nabih is going to show us later. So it will be a training on something new and something, that, hasn't been exposed before. So, just take a look because, this year in Las Vegas, there will be, also a place for for Alfresco developers in the CommunityLIVE. And finally, just, remember that everyone is welcome to the Afrisco community.

Use the Hyland connect for your doubts and join the Hyland community as customers if you want to get the latest release notes and and so on. We have also some events related to our first going to some other, products in the in the company. You can also join our Discord server, obviously, and and just speak with us. And there is the two different projects, the alfresco that is the official repository for the alfresco projects, and also the Africa labs with some experiments that are not officially supported by alfresco as company, but that you can also use. Also, if I miss something in my news and you want to be, here in the nest detail, just contact with me, and I will include also your your updates in this.

And with that, let's move to today's presentation. So it's the, transform your content into AI ready data with knowledge enrichment APIs. So my colleague, Nabih Metri, that is product monitoring Hyland is coming from document filters. So he has deep knowledge in transformation, right? In all these kind of what in

our fresco is done by the transport service and the transport core.

There is a different product that is dock filters, and that is also used as base for this knowledge and treatment API. But better than me, Nabih will will explain you, how this works. So I I just pass the the wall to you, Nabih. So, yes, please, go ahead.

Nabih Metri: Thank you. Thank you. Just give me one second to get this started.

Angel Borroy: Alright.

Nabih Metri: So as Anil said, I'm gonna be I am the product manager over knowledge enrichment along with document filters. So I'm really glad that you're all here today because we have some really exciting stuff to tell you about. We're going to talk about a challenge that touches nearly every organization today. How to take the massive amounts of unstructured content you already have and turn that into something actionable, something intelligent. So our solution is called Knowledge Enrichment.

And it's designed to make unstructured data not only searchable, but contextual, structured, and usable in other systems. So let's start with the big picture. So eighty percent of enterprise data is unstructured. So this includes emails, scanned documents, contracts, images, presentations, and so much more. And this data isn't all in one place.

The average enterprise has twenty one different content repositories. That's everything from on prem SharePoint to cloud storage, content management systems, email archives, and a lot more. So even before we can do AI, we have to first ask, can we even find the data we need? Can we trust it? Can we take action on it?

So let's talk about why that eighty percent stat is so painful. First, there's the increasing number of data sources. Your content lives in too many systems. That makes it hard for employees, let alone AI, to connect the dots. Then there's the lack of content structure.

It's not just that the data is messy. It's that it lacks meaning. And lastly, there's the inability to take action. Fragmented, inconsistent data leads to poor decisions. Imagine a claims system that receives hundreds of PDF documents but can't tell which ones are invoices, repair estimates, or photos.

AI can't act on what it doesn't understand. That's the heart of the problem knowledge enrichment addresses: bad data. So I wanted to give you a real life example of how the quality of data can greatly impact the output of a system. So this is a recipe for a meringue cake. If anyone has followed a recipe before, you can tell that something just isn't right.

The ingredients mention a secret ingredient. The instructions don't mention what temperature to set the oven to. It even mentions whipped cream but didn't give directions for how to make it. This can only lead to disaster. So this broken

cake is what happens to the output that an AI system will give if you use bad data.

Please don't be that person that uses bad data. I wanna see what good data can do. Here we have again a recipe for a meringue cake, but this time, it has a very detailed ingredient list and clear instructions that are detailed enough where they won't even fit on the screen. The amount of detail and context in this recipe is the key to a great outcome. Now, this cake looks so much better than the last one.

And with knowledge enrichment, it can help get your data in the state to create outcomes as delicious as this cake. So what even is knowledge enrichment? At its core, it's about transforming raw, unstructured content into structured, contextualized data ready for use in AI and other systems. So this isn't just about better search or better tags. It's about making your content work like data so it can be analyzed, automated, and trusted in your decision making workflows.

So knowledge enrichment is currently available in beta, and I'll later show you how you can get your hands on it for free. So knowledge enrichment is designed as a set of SaaS based APIs built specifically for those creating AI systems, intelligent workflows, and automation platforms. It's not tied to any other highland product in the sense that we require you to use OnBase or Alfresco or Nuxeo or Perceptive. It can work with all of them. It can even work with anything else that you build since it is just a set of APIs.

So we've split this into two parts: data curation, which turns raw content into AI ready structured data, and then context enrichment is then the piece that adds the surrounding meaning and contextual awareness so systems can make better decisions. So this is about making real world content usable by modern systems. So in the knowledge enrichment data, we're offering both data curation and context enrichment capabilities. So everything here is available right now. So that includes support for over six hundred different file formats, ranging from Microsoft Word, PowerPoint, Excel, to Apple iWorks files, to PDFs, images, and even WordPerfect from the '90s.

Oh, and I can't forget audio video transcription. And this support isn't powered by LLMs. It's primarily powered by Hyland's own document filters, which provides best in class text extraction. It then will structure the extracted text on the source file into Markdown or JSON, so it can be used in your AI system, also powered by document builders. But that's not all for data curation.

It provides the ability to normalize text. It has a trunking strategy to give contextual chunks and even provides positional information around the content. Content enrichment then takes the next step. It surfaces metadata, relationships, and inferred meaning that's critical for AI systems. And these are just the beta features.

We're going to continue improving these features to get more tailored results and

provide additional functionality so you can take even more precise actions with your data. Now let's take a look at what knowledge enrichment can actually do. So here we have a repair estimate for a damaged roof of a house. You can see that it has different sections. It has headings, paragraphs, things are bolded, and even have a table.

Using data curation, we're able to extract the content of that file, convert it to markdown, all while keeping the structure of the document. Or how I've heard it said by the great Ben Truscott, we're keeping the author's intent. We're not converting the source file. We're not using OCR or AI models to extract content from the text based documents. We're pulling the content straight from the file, including the table information using Hyland document filters.

Now, you all may be asking, Okay, Nabi, I mean, that's cool, but does this actually change the output that I'll get? Let's take a look and find out. So I can be a slacker sometimes, and this was one of those moments, which is why this is not a live demo. So I apologize for that. But you can get a trial of, knowledge discovery if you want to try this out for yourself.

So I borrowed this recording from the Hyland Knowledge Discovery product manager, Tyler Khan. Knowledge Discovery is HANA's chat agent that lets you ask questions and get answers about your contacts. But what you'll see here is a comparison of the output you get from Knowledge Discovery using two different data preparation pipelines. The first will be an industry standard, and the second will be Hylin's own data curation. So we'll start out by asking a question that we want to know the answer to.

So you can see that the system wasn't really able to provide an answer. So we'll switch to use data curation and see what we get as that result. Now that's a huge difference. Using an industry standard solution wasn't able to provide meaningful information, but a simple switch to using data curation to provide the system with good data completely changed the result. And the documents that were used to get this answer are shown here in Knowledge Discovery, and we see that they are PDFs.

So it's the cake situation all over again. Now let's see what context enrichment we can do. So we have that same roof repair estimate we saw before. But this time, we're not only structuring the content. We want to learn more about it.

So traditional metadata can be extracted like the document type, company name, and total estimate, but the key to context enrichment is being able to generate enriched metadata. So we can determine things that aren't directly in the document, but we can infer about the content. We can learn more about the property and infer that this estimate is for a residential home. Or we can learn more about the damage and infer that there's significant damage to the roof, potentially from a recent storm. So this level of context can be vital to a workflow or even part of an AI system.

And typically, when there's a roof repair estimate, there are pictures as well. So

contact enrichment can do its magic on images too. It can analyze the image and tell you more information, sometimes even things that a person wouldn't notice. So in this case, it can give the location of the damage, how severe the damage is, what kind of damage there is, and even a potential cause of the damage. Now, again, this on its own is great.

It can help ensure that data is consistent, accurate, and even just made available to the user. But why would we stop here? Why are we looking at each document individually? So we have great data now using knowledge enrichment. We were able to structure it.

We were able to structure an unstructured document and generate insights about it. It. We were able to infer new information from an image and create enriched metadata. Now let's take the next step. Let's actually take action on the output that knowledge enrichment was able to provide.

Using knowledge enrichment, we can create a fraud detection AI agent that correlates the output from the roof repair estimate to the damage shown in the image to indicate if there is suspected fraud. Now, this agent piece isn't available from Hyland yet, but it will be soon. But that doesn't stop you from just using knowledge enrichment to build your own agent, because as long as we can get you good, structured, enriched data, the sky's the limit. Now, after all that talking about knowledge enrichment, what it can do and what it can lead to, it's only fair to ask the question, what makes it different? So knowledge enrichment combines precision extraction with spatial context to create smarter, trusted AI pipelines.

First, we start with extraction quality, powered not by LLMs, but by high end document filters. So this is an important distinction. While LLMs can be impressive, they're not designed for high fidelity content extraction. We are. Our technology performs clean, precise extraction across over six hundred different file formats.

It's not just about grabbing text. It captures tables, formatting, structure, layout, even positioning, all while being lightweight, which brings me to the next point. Structure and position are preserved. We don't give you a blob of text and ask you to figure it out. If a table exists in the document, it comes out as a table.

Headers state headers. Lists state lists. Sections are recognized. That means your outputs are ready to use right out of the box. There's no need for extra cleaning, formatting, or post processing.

That's a huge time saver, and it reduces downstream errors in AI pipelines. And finally, positional intelligence is embedded directly into the output. Whether you're working with JSON or Markdown, our outputs include context about where each element was located on the page. I don't know if I've ever seen that with any other markdown output. So this enables layout aware AI workflows.

You can do table schema mapping. You can associate text with its visual

groupings on the page. You can build agents that don't just read documents, but visually understand them. That's why this isn't just another enrichment tool. It's a foundation for smarter, more reliable AI systems that understands not just of what the content says, but how it's structured, where it belongs, and why it matters.

As I've said before, we're currently in beta and welcoming new users to join. So it's a great time to test the APIs, help us shape the roadmap, and see how knowledge enrichment can help you get more from your unstructured content. So you can scan the QR code or click the link from the slides I'm reading available. I've also included a link to the API documentation so you can get an idea of what's provided prior to joining beta. So as I wrap up and for you, for my voice here are five questions to think about.

Have you identified which business processes rely on unstructured or semi structured data? Can you list the specific decisions or actions that would benefit from enriched, contextualized data? Do you know where your most valuable data lives and in what formats and systems? Have you assessed the quality, completeness, and context of your data before it reaches downstream AI systems? Has your team agreed on how you'll measure the impact of cleaner, more structured knowledge on business outcomes or AI performance?

So these are great starting points to explore if knowledge enrichment is a fit for your organization. So thank you again for your time and technical today. I hope this has sparked some ideas about what's possible when you treat unstructured content not as a problem, but as an untapped asset. So we'd love to talk more, hear about your use cases, and get your feedback in helping us shape Knowledge and Reshma and our other offerings. And with that, I will now pass it off to

Angel Borroy: Okay. So thanks for the presentation. I don't know if you have any any questions from the from the attendees, but I have some questions as an Afresco user. Right? So okay.

So you were talking about this knowledge enrichment, API. So I guess that as a developer, there should be some API I can use in order to send my documents, to the CIC. So can you explain more how that works from an Alfresco developer point of view?

Nabih Metri: Yes. So as an Alfresco developer or any developer in general, the way that using knowledge enrichment would work is we have a set of SaaS based APIs. So they are, at the moment, currently, it is AWS environment that is hosted within the US. Now we're looking to expand that for other other cases, but that's how it is right now. So you would call our API.

You would pass it the document, document, image, whatever file that you want, that you want to pass to knowledge enrichment. And then you would indicate what action you want to take. So if you want just a markdown output of the file, you tell us that, and we'll give that right back to you in markdown. If you want us to chunk your document based off of that markdown, do that too. If

you want embeddings, you can do that.

If you want us to generate new metadata, it is the same process. You just feed us that document, and we feed you the output that you want, and you can use that output wherever you want and however you want.

Angel Borroy: And that that's that's great to know. So, just let me re elaborate. Right? So this is an individual service with the goal of, providing information in relation with your documents. So, I mean, I can pass the documents to knowledge enrichment, and then I can get back the information extracted from the document, the markdown, the whatever, all the information that is extracted from the document.

And then I can create my own on premise AI, vector database, whatever system. Right? So

Nabih Metri: Yes. Exactly.

Angel Borroy: Right. But this is also part of the CIC. So if I'm using the CIC, then I also get this Yes. Okay.

Nabih Metri: That that is a great point. So the other CIC products we have will be using or even currently use knowledge enrichment on the back end. So if you're using, for example, knowledge discovery, it's using knowledge enrichment to do its data curation, to do its context enrichment. So you don't need knowledge enrichment if you are using one of these other products if you just want those other products. But if you're looking to build something with what knowledge enrichment can provide, then knowledge enrichment is what you're looking for.

Angel Borroy: Okay. So so we have also I I will I will, make some more questions later. But we have a question from from, having up. And that is a common suspect in the in the community. Can you talk about the data privacy regulation and compliance?

Okay.

Nabih Metri: So, we are not currently compliant with those frameworks at the moment, but that is something we are working towards. Each CIC product will have its own journey for how it gets there. Knowledge enrichment will be an easier step because we are not storing your data unless you want us to. So we have our own data lake. If you want to utilize that, we can store your, your embeddings in there, but that's not required.

So you could just use knowledge enrichment as a pass through to just process your data, and we don't store anything. But at the moment, we're not officially GDPR or DORA compliant, but that is something we are working towards.

Angel Borroy: Okay. Because I guess that this is not a question for you. But as we have Haim with us, there is some plans, some road map in order to run

this part this module, this service on premise at some point on a on a private, cloud or something like that. Hello, Hain.

Nabih Metri: Hello.

Hein Ragas: So, yes, we are currently thinking of how could we provide a a connector similar to the Elasticsearch connector or even the, connector for knowledge discovery that allows Alfresco to automatically call knowledge enrichment and get the data back and put that data as metadata on on on the document. So that would take away all of the software development that is needed to, you know, to to to call because knowledge enrichment is a set of APIs. Today, you have to do the coding yourself. So if we have something like a connector, yeah, that that automatically based on I don't know. An aspect is set, and that triggers a call to knowledge enrichment, and the metadata comes back and is placed to another metadata field, for the documents, that something like that would take away, the requirement to code it yourself.

And then the metadata is or the, you know, the new data is in alfresco and available available there?

Angel Borroy: Okay. So so this should be, like, the first step. So when we are compliant with all these, kind of regulation, then we will be able to use that from our FreshCo and use that as a real service. So just to transfer my information. But how about the next step?

How about someone just trying to run the full knowledge enrichment service compromise? Is that an option or not?

Hein Ragas: So I think we are exploring that, ways to do hybrid deployments with more and more components from knowledge enrichment on premise. But that is you know, those are still partially baked ideas. It's still early days. You know, knowledge enrichment is is available in beta. You know, it's it's not even officially released yet.

So so certainly we're we're open, you know, we're open, to to discussions with people who have opinions or ideas. Please reach out. But, yes, those you know, we're we're thinking towards that direction. Yes.

Angel Borroy: No. No. That's that's perfectly fine. I mean, it's just to express, that we are thinking on all these options that they are considered. They are not yet on our official road map, but we know that these requirements can be done in the future.

But, also, we need to think that we are using a lot of resources, together with document filters in order to produce high quality, transformation processes. Right? So if you want to replicate this, then you need to first pay a lot of money because you need a lot of resources and then, high complexity in terms of deployment and this kind of office stuff. Right? But Yes.

It's it's fine. So let let's take another question. So I don't know if that is already, reply. So can you can you, Nabih, reply this, for the recording, this one from

Rodrigo? Yeah.

Nabih Metri: So the question was, does knowledge enrichment provide the capability to bring your own model? And at the moment, it does not, but it will soon. So that is something that we know that is very needed by a lot of people. We currently use foundational LLMs to do any, any model work that we have to do, but we know that that isn't always the best case. So either for regulatory reasons, you want to use your own, or it just may not give the best result.

So we're looking to provide the ability for you to bring your own models to to choose what models you wanna use. So if you have your own credentials to say OpenAI, you'll be able to use those. Or we even want to be able to just use different models, like a small language model, maybe one that is specific for your use case. So if you're insurance, maybe a model that is good for insurance that can provide better results using that. And then I believe the next question was around pricing.

So, yes, pricing is available. It is available by request since we're still refining it. But we it is primarily usage based and definitely open for feedback. This is one of the first products that we are providing like this. So we're pricing it in the way that we believe is best.

But if you don't believe is that that that is best, we want to know that because we want to provide something that is valuable for you.

Angel Borroy: Okay. Perfect. So there is a new question. So that's I I I don't know, to be honest. So that's the CIC IDP use k e k e.

I don't know what's No.

Nabih Metri: I was in Richmond.

Angel Borroy: Oh, okay. No. That's not good. Okay. You're doing now just using acronyms, so that's great.

Okay. Please go ahead.

Nabih Metri: Yeah. So we are actually working with the IDP team to switch it over to be to use knowledge enrichment and to expand its capabilities.

Angel Borroy: Okay. Perfect. So great. I don't know if there is any other, there is a new one. Does it support multilingual documents?

I guess that the answer is yes.

Nabih Metri: Yes. Yes. So any type of document, regardless of format or language, is supported. So we don't we don't discriminate. Any kind of language you want the document to be in, we'll be able to work with.

Angel Borroy: Okay. Perfect. That that makes makes sense. Which model is used now with LLM, I guess, that is the one used right now?

Nabih Metri: So we are currently exposing a few different models for different use cases. So it just depends on what your what functionality you're looking to to use. Well, they are all, foundational LLMs at the moment.

Angel Borroy: And and as I guess that they are all big LLMs. Right?

Nabih Metri: Yes.

Angel Borroy: Okay. So state of the art. LLMs, depending on if you want just visual encoding or or not, then we need to choose between one or the other. Okay. I don't know if there is more question.

Okay. Would this be an additional cost?

Nabih Metri: It would not. So this is baked into Niles discovery. So, like, when you're buying for example, if you're buying a car, you're not paying for a seat in the car. You're paying for the car. That's kind of how knowledge discovery is being is being approached.

You're paying for knowledge discovery, and it just happens to use knowledge enrichment on the back end to help power it.

Angel Borroy: Okay. Great. So the answer is no. So I don't know if there is more questions. So I'm just so my final thought should be that this is really a key service in every, document repository.

So if you are able to get the most of your documents, then you will get, the most of your AI services. But in any case, Rodrigo is, again, asking about OCR. So OCR requires to be exact and images in good quality.

Nabih Metri: So when we need to OCR a document, it is because it's an image based text document. You can use any kind of document, good quality or not. With the good quality, you will get better results. And then we are looking to even improve our current OCR offering. So we're using a partner OCR service at the moment.

We're looking to just see what can we do better to get better results out of bad data.

Angel Borroy: And, also, we have Hyland. It has also some other products, to perform this OCR with, like, the best quality in the market. So if that is not enough, then you can even just make some preprocessing round before sending the the information to knowledge, discovery. So, again, I don't know. Not if you want just to, say something else, a closing remark, but it looks like they are happy with it.

Nabih Metri: Yeah. So, definitely check out the API documentation. It's in the slides, and I put a link in the chat as well. That'll give you an idea for how you can start using the API if you decide to be a beta customer. If you're looking for other CIC products, knowledge enrichment will be used on the back end, so you don't even need to worry about it at that point.

But if you want to build those mind blowing AI applications, if you want to build your own AI agents, knowledge enrichment is going to be key.

Angel Borroy: Yeah. Of course. So this will increase, with in one click the quality of your results with with AI. So it was really amazing to to have you both today. We will be just, publishing the recording, and you can connect with us for any further questions via this, connect blog post.

We hope to see you in in Las Vegas next August because we will have news also about this, knowledge enrichment model. So thank you to everyone, till the next time. Thanks. Bye.

Thank you. Bye. Thanks, Hein.